

00-27-2

PTO/SB/21 (09-04)

Approved for use through 07/31/2006. OMB 0651-0031

U.S. Patent and Trademark Office: U.S. DEPARTMENT OF COMMERCE

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

AF/1637
JHW

| | | | |
|--|----------------------|------------------------|--------|
| O I P E TRANSMITTAL FORM JUN 26 2005 (to be used for correspondence after initial filing) | Application Number | 09/827,383 | |
| | Filing Date | April 4, 2001 | |
| | First Named Inventor | Michael Mittmann | |
| | Art Unit | 1637 | |
| | Examiner Name | Jeffrey Norman Fredman | |
| Total Number of Pages in This Submission | 40 | Attorney Docket Number | 3108.1 |

| ENCLOSURES (check all that apply) | | |
|---|--|---|
| <input checked="" type="checkbox"/> Fee Transmittal Form (in duplicate) <input type="checkbox"/> Fee Attached <input type="checkbox"/> Amendment / Reply <input type="checkbox"/> After Final <input type="checkbox"/> Affidavits/declaration(s) <input type="checkbox"/> Extension of Time Request <input type="checkbox"/> Express Abandonment Request <input type="checkbox"/> Information Disclosure Statement <input type="checkbox"/> Certified Copy of Priority Document(s) <input type="checkbox"/> Reply to Missing Parts/Incomplete Application <input type="checkbox"/> Reply to Missing Parts under 37 CFR 1.52 or 1.53 | <input type="checkbox"/> Drawing(s) <input type="checkbox"/> Licensing-related Papers <input type="checkbox"/> Petition <input type="checkbox"/> Petition to Convert to a Provisional Application <input type="checkbox"/> Power of Attorney, Revocation Change of Correspondence Address <input type="checkbox"/> Terminal Disclaimer <input type="checkbox"/> Request for Refund <input type="checkbox"/> CD, Number of CD(s) _____ <input type="checkbox"/> Landscape Table on CD | <input type="checkbox"/> After Allowance Communication to TC <input type="checkbox"/> Appeal Communication to Board of Appeals and Interferences <input checked="" type="checkbox"/> Appeal Communication to TC (Appeal Brief - 16 pages) <input type="checkbox"/> Proprietary Information <input type="checkbox"/> Status Letter <input checked="" type="checkbox"/> Other Enclosure(s) (please identify below): - Exhibits 1, 2 and 3. - Return postcard |
| Remarks | | |

| SIGNATURE OF APPLICANT, ATTORNEY, OR AGENT | | | |
|--|------------------------|----------|--------|
| Firm | Altymetrix, Inc. | | |
| Signature | <i>Sandra E. Wells</i> | | |
| Printed Name | Sandra E. Wells | | |
| Date | June 24, 2005 | Reg. No. | 52,349 |

| CERTIFICATE OF TRANSMISSION/MAILING | | | |
|---|---------------------|------|---------------|
| I hereby certify that this correspondence is being deposited with the United States Postal Service with sufficient postage as EXPRESS MAIL post Office to Addressee in an envelope addressed to: Mail Stop: APPEAL BRIEF, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450 on the date shown below. EXPRESS MAIL LABEL: EV 500852262 US. | | | |
| Signature | <i>Pamela Leong</i> | | |
| Typed or printed name | Pamela Leong | Date | June 24, 2005 |

This collection of information is required by 37 CFR 1.5. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to 12 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

Effective on 12/08/2004.

Fees pursuant to the Consolidated Appropriations Act, 2005 (H.R. 4818).

**FEE TRANSMITTAL
for FY 2005**☐ Applicant claims small entity status. See 37 CFR 1.27**TOTAL AMOUNT OF PAYMENT** (\$) 500.00**Complete If Known**

| | |
|----------------------|------------------------|
| Application Number | 09/827,383 |
| Filing Date | April 4, 2001 |
| First Named Inventor | Michael Mittman |
| Examiner Name | Jeffrey Norman Fredman |
| Art Unit | 1637 |
| Attorney Docket No. | 3108.1 |

METHOD OF PAYMENT (check all that apply)☐ Check ☐ Credit Card ☐ Money Order ☐ None ☐ Other (please identify) : _____☒ Deposit Account Deposit Account Number: 01-0431 Deposit Account Name: Affymetrix, Inc.

For the above-identified deposit account, the Director is hereby authorized to: (check all that apply)

☒ Charge fee(s) indicated below☐ Charge fee(s) indicated below, except for the filing fee☒ Charge any additional fee(s) or underpayments of fee(s)☒ Credit any overpayments

Under 37 CFR 1.16 and 1.17

WARNING: Information on this form may become public. Credit card information should not be included on this form. Provide credit card information and authorization on PTO-2038.**FEE CALCULATION****1. BASIC FILING, SEARCH, AND EXAMINATION FEES**

| Application Type | FILING FEES | | SEARCH FEES | | EXAMINATION FEES | | Fees Paid (\$) |
|------------------|-------------|-----------------------|-------------|-----------------------|------------------|-----------------------|----------------|
| | Fee (\$) | Small Entity Fee (\$) | Fee (\$) | Small Entity Fee (\$) | Fee (\$) | Small Entity Fee (\$) | |
| Utility | 300 | 150 | 500 | 250 | 200 | 100 | _____ |
| Design | 200 | 100 | 100 | 50 | 130 | 65 | _____ |
| Plant | 200 | 100 | 300 | 150 | 160 | 80 | _____ |
| Reissue | 300 | 150 | 500 | 250 | 600 | 300 | _____ |
| Provisional | 200 | 100 | 0 | 0 | 0 | 0 | _____ |

2. EXCESS CLAIM FEES**Fee Description**

Each claim over 20 (including Reissues)

Each independent claim over 30 (including Reissues)

Multiple dependent claims

| | | | |
|---------------------|---------------------|----------------|----------------------|
| Total Claims | Extra Claims | Fee(\$) | Fee Paid (\$) |
|---------------------|---------------------|----------------|----------------------|

_____ - 20 or HP = _____ x _____ = _____

HP = highest number of total claims paid for, if greater than 20.

| | | | |
|----------------------|---------------------|----------------|----------------------|
| Indep. Claims | Extra Claims | Fee(\$) | Fee Paid (\$) |
|----------------------|---------------------|----------------|----------------------|

_____ - 3 or HP = _____ x _____ = _____

HP = highest number of independent claims paid for, if greater than 3.

| Small Entity | |
|---------------------------|---------------|
| Fee (\$) | Fee (\$) |
| 50 | 25 |
| 200 | 100 |
| 360 | 180 |
| Multiple Dependent Claims | |
| Fee (\$) | Fee Paid (\$) |
| _____ | _____ |

3. APPLICATION SIZE FEE

If the specification and drawings exceed 100 sheets of paper (excluding electronically filed sequence or computer listings under 37 CFR 1.52(e)), the application size fee due is \$250 (\$125 for small entity) for each additional 50 sheets or fraction thereof. See 35 U.S.C. 41(a)(1)(G) and 37 CFR 1.16(s).

| | | | | |
|---------------------|---------------------|---|-----------------|----------------------|
| Total Sheets | Extra Sheets | Number of each additional 50 or fraction thereof | Fee (\$) | Fee Paid (\$) |
|---------------------|---------------------|---|-----------------|----------------------|

_____ - 100 = _____ / 50 = _____ (round up to a whole number) x _____ = _____

4. OTHER FEE(S)

Non-English Specification, \$130 fee (no small entity discount)

Other (e.g., late filing surcharge): Appeal brief fee under 37 CFR § 41.20(b)(2)

| Fees Paid (\$) |
|----------------|
| 500.00 |

SUBMITTED BY

| | | | | | |
|-------------------|------------------------|--------------------------------------|---------------|-----------|--------------|
| Signature | <i>Sandra E. Wells</i> | Registration No. (Attorney/Agent) | 52,349 | Telephone | 408-731-5000 |
| Name (Print/Type) | Sandra E. Wells | Date | June 24, 2005 | | |

This collection of information is required by 37 CFR 1.136. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 30 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.

If you need assistance in completing this form, call 1-800-PTO-9199 (1-800-786-9199) and select option 2.



**IN THE UNITED STATES PATENT AND TRADEMARK OFFICE
BEFORE THE BOARD OF PATENT APPEALS AND INTERFERENCES**

In re application of:

Mittmann et al.

Serial No: 09/827,383

Filing Date: April 4, 2001

Title: TAG NUCLEIC ACIDS
AND PROBE ARRAYS

Examiner: J. Fredman

Group Art Unit: 1637

Mail Stop Appeal Brief-Patents
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

APPEAL BRIEF

Sir:

Applicants submit this Appeal Brief in response to the final rejection mailed January 31, 2005 and pursuant to the Notice of Appeal filed on April 27, 2005.

I. REAL PARTY IN INTEREST

The real party in interest in this application is Affymetrix, Inc.

II. RELATED APPEALS AND INTERFERENCES

Based on information and belief, there are no related appeals or interferences.

III. STATUS OF CLAIMS

Claims 3-6, and 8-14 are canceled. Claims 1, 2, 7 and 15-19 stand rejected as lacking utility and for lack of enablement based on the finding of lack of utility. Claims 1, 2, 7 and 15-19 are appealed.

IV. STATUS OF AMENDMENTS

No amendment after final rejection has been filed. The appealed claims are shown in Appendix I.

V. SUMMARY OF CLAIMED SUBJECT MATTER

Independent claim 1 is drawn to a set of at least 1,000 nucleic acid tag probes selected from SEQ ID NOS: 1-2000. Support for this claim may be found in originally filed claim 4.

Independent claim 2 is drawn to a set of at least 1,000 nucleic acid tags selected from the complements of SEQ ID NOS: 1-2000. Support for this claim may be found in originally filed claim 4.

Independent claim 16 is drawn to a set of nucleic acids including each of the sequences of SEQ ID NOS: 1-2000, the complements of SEQ ID NOS: 1-2000, each of the sequences of SEQ ID NOS: 1-2000 with a mismatch at the 10th nucleotide and the complements of the mismatches to SEQ ID NOS: 1-2000. Support for this claim may be found on page 8, line 28 to page 9, lines 1 to 4 and in originally filed claim 12.

SEQ ID NOS: 1-2000 are included in the sequence listing and in Table 1, pp. 10-53. The sequences are each 20 nucleotides in length and they were selected from a list of all possible 20 mers and they do not correspond to any known genomic sequences. The sequences are not expressed sequence tags (ESTs) and were selected for inclusion in the set because they do not occur naturally in the genome of an organism. The sequences were selected for inclusion in the set if they met a series of filtering criteria designed to optimize and standardize the hybridization characteristics of the set, for example, all

members of the set have closely matched melting temperatures and no member of the set is identical or nearly identical to another member of the set or to a sequence in the public database (page 8, lines 20-25). A filtered set of 2200 tag sequences was synthesized and hybridized to an array containing the complementary 2200 tag probes and 2050 pairs showing the highest discrimination and signal intensity were selected for inclusion in the application (page 9, line 5-7). The sets of nucleic acids may be attached to a solid support (Page 3, lines 2-10). The sets may include controls including probes with a mismatch at position 10 of the probe (Page 53, lines 34-37).

VI. GROUNDS OF REJECTION TO BE REVIEWED ON APPEAL

Claims 1, 2, 7 and 15-19 stand rejected under 35 U.S.C. 101 as lacking patentable utility and under 35 U.S.C. 112 as failing to comply with the enablement requirement.

VII. ARGUMENTS

The claimed invention has patentable utility and meets the requirements of 35 U.S.C. 101.

Section 101 of Title 35 of the United States Code states that for an invention to be patentable it must be useful:

Whoever invents or discovers any new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof, may obtain a patent therefore, subject to the conditions and requirements of this title.

Following the requirements of the Utility Examination Guidelines published at 66 FR 1092, Jan. 5, 2001, superseding the Revised Interim Utility Examination Guidelines that were published at 64 FR 71440, Dec. 21, 1999; 1231 O.G. 136 (2000); and correction at 65 FR 3425, Jan. 21, 2000; 1231 O.G. 67 (2000), a rejection based on lack of utility should not be imposed if the claimed invention has either a (1) well-established utility or the applicant has (2) asserted a specific and substantial utility that is credible. An assertion that the claimed invention is useful for a particular purpose is sufficient provided that the assertion would be considered credible by a person of ordinary skill in the art.

As indicated on pages 2 and 3 of the office action mailed January 31, 2005 the Examiner indicates that the specification asserts at least one credible utility: “to analyze genomic DNA”, but asserts that the claimed invention lacks a utility that is substantial and specific. The Examiner appears to be incorrectly equating the claimed set of tag sequences with ESTs or genomic sequences. On page 3 of the office action the Examiner indicates that “if there were evidence of the association of SEQ ID NO: 1 with any disease state, with a protein activity or with some other biological phenotype, this evidence might be considered regarding a substantial utility.” On page 4 the Examiner states that “the instant claims are drawn to a set of polynucleotides with no known function” and that the “function of the nucleotides is as yet undetermined with no known biological significance”. This reasoning appears to ignore the fundamental nature of tag sequences.

If the claimed sequences were derived from genomic sequence such an association would be required (see Ex parte Fisher 72 USPQ2d 1020 (2004)), but the claimed sequences were specifically chosen because they were not genomic sequences and were not expected to have a biological function. They do not code for protein and are not part of a naturally occurring gene. If the sequences were found to have association with a disease state or a biological phenotype they would not be useful for any of the disclosed utilities of tags.

As set forth in Brenner v. Manson, 383 U.S. 519, 148 USPQ 689 (1966), 35 U.S.C. § 101 requires a utility that is “substantial” i.e., one that provides a specific benefit to the public in currently available form. Brenner, 383 U.S. at 534-35, 148 USPQ at 695. The threshold of utility under this standard is not high, and is met if an invention is capable of providing some identifiable benefit. Juicy Whip Inc. v. Orange Band Inc., 185 F. 3d 1364 at 1366, 51 USPQ2d 1700 (Fed. Cir.1999). As set forth in the Revised Interim Utility Guidelines Training Materials at pages 5 and 6, a “specific utility” is a utility that is specific to the subject matter claimed in contrast to a general utility that would be applicable to the broad class of the invention and a “substantial utility” is a utility that defines a “real world” use.

Applicants have asserted several specific and substantial utilities for the claimed invention and the invention provides a specific benefit to the public in its currently available form.

In the specification on page 2, lines 12-18, the use of “tags” to identify specific biological substance in a sample is disclosed along with several background references providing examples of methods where sets of tags, such as those presently claimed,

would be useful (page 2, lines 19-22). For example, US Patent No. 5,541,505 (page 2, line 18) discloses methods for tagging materials, such as explosives, with nucleic acids and tracing the materials by detection of the nucleic acid tags. The specification of the '505 patent discloses a method for tagging a material by attaching a nucleic acid tag of specific sequence to the substance. The presence of the tag can be detected to provide a method to trace the tagged material (column 1, lines 50-60). Substances that may be tagged in this manner include oils, foodstuffs, explosives, medicaments and paint products (column 2, lines 10-14). The '505 patent also discloses that sequences that are useful for the disclosed tagging methods are "typically non-biologically functioning in that they are not a [part] of a functional nucleic acid sequence operating in a living cell" (column 2, lines 62-65). The sets of tags presently claimed may be used for the methods of tagging and tracing materials disclosed and claimed in the '505 patent.

The specification also indicates that the claimed sets of tags would be useful in the methods disclosed in US Patent No. 5,654,413 (page 2, line 18). The '413 patent discloses uses for tags including as a mechanism for identifying electrophoretically separated bands on a gel that consist of DNA fragments generated in the same sequencing reaction and for "manipulating and identifying individual molecules in complex combinatorial chemical libraries, for example, as an aid to screening such libraries for drug candidates" (column 1, lines 35-56). The '413 patent also states that "the successful implementation of such tagging schemes depends in large part on the success in achieving specific hybridization between a tag and its complementary probe" (column 1, lines 61-63) and that a repertoire of tags that minimize occurrence of false positive and

false negative signals (cross-hybridization), such as the tags presently claimed, would be useful for such tagging methods (column 2, lines 45-49).

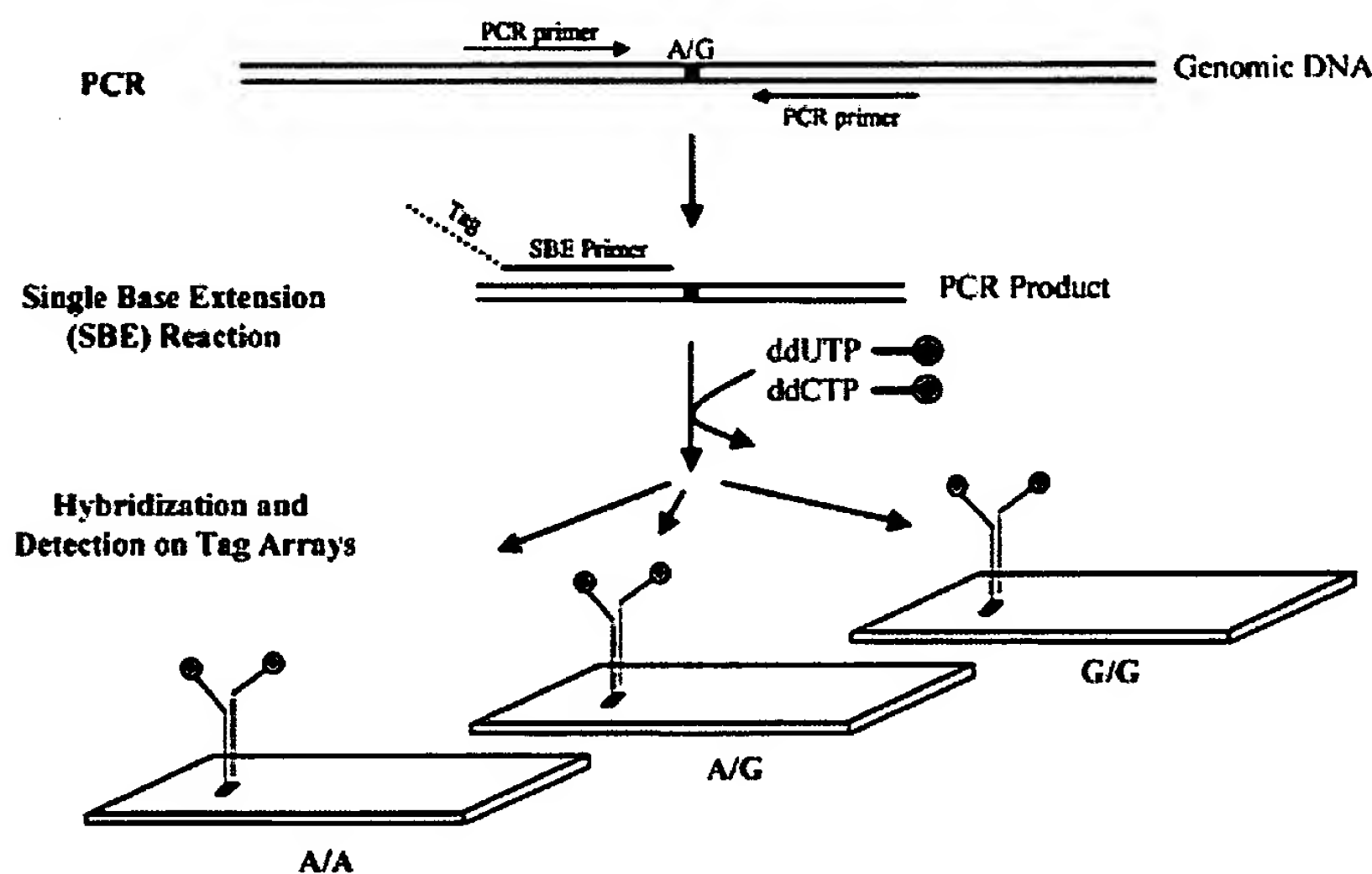
The specification also provides detailed disclosure of methods of use for the claimed set of tags. On page 10, lines 15-22 the use of tagged primers and an array of tag probes for genotyping polymorphisms is disclosed as follows:

One embodiment of the method involves three step[s]: (1) amplification of the polymorphic locus, (2) primer extension of a sequence-tagged primer with distinct labels for different polynucleotides at the polymorphic locus, and (3) hybridization to a tag array. The amount of each distinct label can be determined at known positions of the tag array. Each tag represents a distinct polymorphic locus and each distinct label represents a distinct allelic form at the polymorphic locus. The method permits the simultaneous determination of a genotype at multiple loci, as well as the determination of allele frequencies in a population. Another embodiment employs just steps (2) and (3).

A schematic of this embodiment is shown in Figure 1 below (see also, Fig. 1 of US Patent Application No. 60/140,359). In this embodiment, SEQ ID NO: 1-2000 are arrayed on a solid support as tag probes and the complements of SEQ ID NO: 1-2000 are used as the tag portion of the primers. A researcher selects a set of polymorphisms to analyze and prepares a tagged-primer for each polymorphic locus to be analyzed. Each primer has a 5' tag region containing the complement of one of the sequences from SEQ ID NOs: 1-2000 and a 3' SBE primer region that is complementary to the region immediately upstream of the polymorphism to be analyzed. Each primer has a known tag sequence and targets a known polymorphism. The set of tagged-primers are hybridized to the fragments containing the polymorphic loci and are extended by a single base corresponding to the polymorphic base. Each species of dNTP present in the reaction is differentially labeled. The labeled tagged-primers are then hybridized to the array of tag

probes and the hybridization pattern is analyzed. The tag probes are present at known locations so the genotype of each polymorphic locus may be determined by determining which label or labels have hybridized to the corresponding tag probe. This method was described in greater detail in US Patent Application No. 60/140,359, which is discussed in the present application at page 10, line 10 and was used in Fan *et al. Genome Res.* 10:853-860 (2000), a copy of which has been provided herewith (Exhibit 1), and was previously provided for the Examiner's consideration with the amendment filed March 4, 2004.

Figure 1. (reproduced from Fan et al. 2004)



Tags are used as a unique label for a particular SBE primer. The tags have uniform hybridization properties so many SNPs can be analyzed simultaneously in a single reaction and detected using the same tag probe array. Instead of designing and making many different arrays with allele specific probes, a single array of tag probes can be used to interrogate many different panels of SNPs, including SNPs that have known

association with disease. As stated in the Revised Interim Utility Guidelines Training Materials on page 6, “an assay that measures the presence of a material which has a stated correlation to a predisposition to the onset of a particular disease condition would also define a “real world” context of use.” Fan *et al.* describes the use of the claimed tags and tag probes for the genotyping of 44 individuals for 142 human SNPs identified in 62 genes that are associated with hypertension. SNPs in promoter regions, at splice junctions and SNPs that alter protein sequence were selected from a larger set of 874 SNPs. Clearly, genotyping SNPs associated with hypertension is a substantial “real world” utility for the claimed invention.

A second utility asserted in the specification at page 10, lines 4-9, is the use of the tags for correlating specific mutations with phenotypic effects. The disclosed tags and tag probes may be used to identify the function of identified open reading frames (ORFs). Deletion mutant strains that are each tagged with a different tag sequence are generated for each ORF, deleting a single ORF in each strain, and the resulting deletion mutants are grown under a wide variety of selective conditions. Detection of the tag sequence indicates survival of the deletion mutant under the selective conditions. This utility of the claimed invention is also described in U.S. Patent Application No. 08/626,285 (now U.S. Patent No. 6,458,530) which is incorporated by reference on page 10, line 5 of the specification, and is also described in Shoemaker *et al. Nat. Genet.* 14:450-456 (1996), a copy of which is provided herewith (Exhibit 2) and was previously provided for the Examiner’s consideration with the amendment filed March 4, 2004.

Shoemaker *et al.* used tags to monitor a population of yeast cells carrying auxotrophic deletions marked with tag sequences. Equal numbers of cells from each of

11 tagged deletion strains were mixed to generate a deletion pool. Aliquots of the deletion pool were grown on complete medium or medium lacking adenine. Tags were isolated from the pools at various time points after inoculation and analyzed on an array of tag probes to determine relative abundance of the different deletion strains. A depletion of the tags associated with the adenine depletion strains was observed after growth on medium lacking adenine but not after growth on the complete medium. Demonstrating that tagging of deletion strains may be used to identify deletions that are responsible for specific phenotypes.

In a more recent study, the claimed sets of tags and tag probes were used to perform multiplex genotyping of 1,121 human single nucleotide polymorphisms from the SNP consortium (TSC) database. Hardenbol *et al.*, *Nat. Biotechnol.*, 21:673-8 (2003), a copy of which has been provided herewith (Exhibit 3). The selected SNPs were from a 16-megabase region on chromosome 6 centered on the linkage peak for IgA nephropathy. A different tag sequence from the set was included in each of 1,121 molecular inversion probes (MIPs), each MIP being specific for a different SNP. The tag sequences in the MIPs were flanked by a 5' locus specific region and a 3' locus specific region (H1 and H2 in Fig. 2a of Hardenbol *et al.*). The MIPs also included universal priming sequences (P1 and P2) and two cleavage sites (X1 and X2). The MIPs are circularized in allele specific reactions and circularized MIPs are amplified by PCR. The genotype of a SNP is determined by hybridization of the amplification products to an array of tag probes that are complementary to the tags in the MIPs. Clearly these publications demonstrate a substantial "real world" use for sets of tags and tag probes.

On page 5 of the Office Action the Examiner indicates that the asserted utilities are not specific to SEQ ID NO: 1-10 because “literally any sequence would function in a genome analysis assay as described in the specification.” Applicants respectfully disagree. The asserted utilities are specific to the subject matter claimed and would not be applicable to the broad class of the invention encompassed by any sequence.

The set of sequences claimed was selected from a larger set of 20 mers by applying a series of filters that included eliminating sequences that could be found in the public databases with fewer than 3 mismatches. There are 4^{20} or $\sim 10^{12}$ different 20 mers that could be made and many different possible combinations of 1,000 or more 20 mers, but only a limited subset of 20 mers met the criteria used for selection of the claimed tags and tag probes. Of the approximately 16,000 starting sequences (pre-selected for similar melting temperatures and GC content) only 2200 met the requirements of all of the filters. Of those 2200, analysis of hybridization intensity and ability to discriminate between perfect match and mismatch during hybridization identified the 2050 probes with the most uniform hybridization behavior. There are other sequences that could be selected as tags and other sets of tags that could be selected, but utility as a tag certainly is not applicable to any sequence or even any 20 mer, for example, no 20 mer that is present in the human genome would be suitable as a tag.

The requirement for a utility that is specific does not require an invention to have a unique utility; a utility may be shared by other compounds or compositions. For example, ibuprofen is useful as an analgesic and has patentable utility even though aspirin and acetaminophen are also useful as analgesics. In contrast, ibuprofen with a disclosed utility to fill a jar that would be useful as a paperweight would not satisfy §101

because the utility is shared by a large class of inventions. Ex parte Fisher at 1028. In this example there is no characteristic of ibuprofen that makes it specifically suited to perform as a paperweight; anything with mass would have the same utility. Unlike the use of ibuprofen as a paper weight in the example, the sequences of the presently claimed invention are selected to have specific, distinguishing characteristics that make them particularly suited to function together as a set of tag probes that may be used to mark genomic sequences in a complex background with a detectable marker sequence. These characteristics are clearly disclosed in the application, unlike the plastic-like invention of In re Ziegler, 992 F.2d 1197, 26 USPQ2d 1600 (Fed. Cir. 1993).

The patentability of the claims is also consistent with prior practice of the US Patent Office in issuing claims to sets of tag sequences. For example, claims 3-10 of US Patent No. 6,458,530 claim sets of at least 100 tag nucleic acids that each have a sequence that meets a selected set of criteria specified in each claim. Claim 8 of the '530 patent is reproduced below:

8. A set of at least 100 different tag nucleic acids wherein:

(a) no two members of the set can be aligned with less than two errors, wherein an error is either a mismatch hybridization or an overhanging nucleotide;

(b) the total number of nucleotides in each tag nucleic acid sequence is identical;

(c) the number of G+C nucleotides in each tag nucleic acid sequence is identical;

(d) the overall number of A+G nucleotides in each of the of the tag nucleic acid sequences is even; and

(e) each member of the set of tag nucleic acids is comprised of a single nucleic acid sequence from the set of nucleic acid tags.

Sets of tags are also claimed in US Patent No. 6,172,214. Claim 1 of the '214 patent is reproduced below:

1. A repertoire of oligonucleotide tags, the repertoire comprising oligonucleotides selected from the group consisting of oligonucleotides of the formula:

$S_1 S_2 S_3 \dots S_n$

wherein each of S_1 through S_n are subunits consisting of an oligonucleotide having a length from three to nine nucleotides and being selected from a minimally cross-hybridizing set;

n is in the range of from 4 to 10, with the proviso that the oligonucleotide tags of the repertoire have a length in the range of from 12 to 60 nucleotides or base pairs; and

wherein there are at least a hundred oligonucleotide tags in the repertoire.

The term "repertoire" is defined as "the set of minimally cross-hybridizing set of oligonucleotides that make up the tags in a particular embodiment or the corresponding set of tag complements" (column 4, lines 30 to 33 of '214). Both of these patents claim unspecified sets of tag sequences that are defined by a set of parameters for sequence selection. The present claims have a similar utility but claim a specific set of tags of defined sequence.

The specification asserts a credible, substantial and specific utility for the claimed invention, making the rejection of the claims under 35 U.S.C. §101 improper. The utility of the claimed invention was disclosed in the specification and has been confirmed by the use of the invention in studies published in peer reviewed journal articles.


The claims comply with the enablement requirement of 35 U.S.C. 112.

The rejection of the claims under the enablement provision of 35 U.S.C. §112 is a corollary of the finding of lack of utility and Applicants request that it be reversed for the same reasons set forth in Applicants' arguments above regarding the rejection under 35 U.S.C. § 101.

For the above reasons, Applicants respectfully submit that rejection of pending Claims 1, 2, 7 and 15-19 is improper. Accordingly, Applicants request that the rejection of Claims 1, 2, 7 and 15-19 be reversed.

Respectfully submitted,

Date: June 24, 2005

By 
Sandra Wells
Reg. No. 52,349

Attachments:

Claims Appendix
Exhibits 1-3

Legal Department
Affymetrix, Inc.
3380 Central Expressway
Santa Clara, CA 95051
Tel: 408/731-5000
Fax: 408/731-5392

Claims Appendix

Listing of Claims:

Claim 1. A set of nucleic acid tag probes comprising at least 1000 nucleic acid sequences chosen from the group consisting of:

SEQ ID NOS: 1-2000.

Claim 2. A set of nucleic acid tags comprising at least 1000 nucleic acid sequences chosen from the group consisting of:

the complements of SEQ ID NOS: 1-2000.

Claims 3-6. (canceled)

Claim 7. The set of nucleic acid tag-probes of claim 1 attached to a solid support.

Claims 8-14. (canceled)

Claim 15 The set of nucleic acid tags of claim 2 attached to a solid support.

Claim 16 A set of nucleic acids comprising:

a first set of nucleic acids comprising each of the sequences listed in SEQ ID NO. 1-2000,

a second set of nucleic acids comprising the complement of each of the first set of nucleic acids,

a third set of nucleic acids comprising the first set of nucleic acids wherein position 10 in each of SEQ ID NO: 1-2000 is changed to a mismatch, and

a fourth set of nucleic acids comprising the complement of each of the third set of nucleic acids.

Claim 17 The set of nucleic acids of claim 16 attached to a solid support.

Claim 18. The set of nucleic acids of claim 16 further comprising a first set of nucleic acids comprising each of the sequences listed in SEQ ID Nos. 2001-2050.

Claim 19. The set of nucleic acids of claim 18 attached to a solid support.

Parallel Genotyping of Human SNPs Using Generic High-density Oligonucleotide Tag Arrays

Jian-Bing Fan,^{1,3} Xiaoqiong Chen,¹ Marc K. Halushka,² Anthony Berno,¹
Xiaohua Huang,^{1,4} Thomas Ryder,¹ Robert J. Lipshutz,^{1,6} David J. Lockhart,^{1,5}
and Aravinda Chakravarti²

¹Affymetrix, Inc., California 95051 USA; ²Department of Genetics and Center for Human Genetics, Case Western Reserve University School of Medicine and University Hospitals of Cleveland, Cleveland, Ohio 44106 USA

Large scale human genetic studies require technologies for generating millions of genotypes with relative ease but also at a reasonable cost and with high accuracy. We describe a highly parallel method for genotyping single nucleotide polymorphisms (SNPs), using generic high-density oligonucleotide arrays that contain thousands of preselected 20-mer oligonucleotide tags. First, marker-specific primers are used in PCR amplifications of genomic regions containing SNPs. Second, the amplification products are used as templates in single base extension (SBE) reactions using chimeric primers with 3' complementarity to the specific SNP loci and 5' complementarity to specific probes, or tags, synthesized on the array. The SBE primers, terminating one base before the polymorphic site, are extended in the presence of labeled dideoxy NTPs, using a different label for each of the two SNP alleles, and hybridized to the tag array. Third, genotypes are deduced from the fluorescence intensity ratio of the two colors. This approach takes advantage of multiplexed sample preparation, hybridization, and analysis at each stage. We illustrate and test this method by genotyping 44 individuals for 142 human SNPs identified previously in 62 candidate hypertension genes. Because the hybridization results are quantitative, this method can also be used for allele-frequency estimation in pooled DNA samples.

The Human Genome Project and other private efforts are producing large amounts of genome sequence and polymorphism data that will provide scientists with an unprecedented opportunity to probe the structure and function of the human genome (Collins et al. 1998). In the realm of human disease, these genomic resources will allow the dissection of the genetic components and molecular mechanisms of complex human diseases and traits. Identification of complex disease genes will require both linkage and association analyses of thousands of polymorphisms across the human genome in thousands of individuals (Risch and Merikangas 1996; Collins et al. 1997; Chakravarti 1999). To enable such large-scale polymorphism analysis in human studies, parallel and efficient genotyping methods are critically needed. The most common variant in the human genome is the single nucleotide polymorphism (SNP) (Wang et al. 1998; Cargill et al. 1999; Halushka et al. 1999). Homogenous and microarray-based minisequencing has been used to genotype SNPs in human populations (Syvanen et al. 1990; Kuppaswamy et al. 1991; Chen and Kwok 1997; Pastinen et al. 1997, 1998; Syvanen 1998). We present a parallel genotyping method for SNPs, termed TAG-SBE, which analyzes al-

lele-specific single base extension (SBE) reactions on standardized, generic high-density oligonucleotide probe arrays (Chee et al. 1996; Shoemaker et al. 1996; Wang et al. 1998; Lipshutz et al. 1999). In TAG-SBE, the array is independent of the specific markers genotyped and the assay can be customized for sets of markers through PCR and SBE primer selection. Because this genotyping method is generic, intrinsically parallel, and favors multiplexed reactions, TAG-SBE is well-suited for large-scale human genetic studies.

To design the tag arrays, all possible 20 mers (4^{20} or $\sim 10^{12}$) were subjected to a computational screen that favored a subset of sequences with similar GC content and thermodynamic properties, and eliminated sequences with possible secondary structure or sequence similarity to other tags (Shoemaker et al. 1996; Giaever et al. 1999; Winzeler et al. 1999). A set of 32,000 tags was selected, with all tags expected to have similar hybridization characteristics and minimal cross-hybridization under standard hybridization conditions. As a hybridization control, and to enable background and cross-hybridization subtraction, each tag probe (PM, perfect match) is paired with a second probe that is identical in sequence except for a single base difference at the central position (MM, mismatch). The high-density tag array used in this study consists of over 64,000 distinct probes, over 32,000 PM tag probes, and over 32,000 adjacent MM probes, each probe occupying an area of $30 \times 30 \mu\text{m}$.

Present addresses: ³Illumina, Inc., San Diego, California 92121 USA; ⁴Kiva Genetics, Inc., Mountain View, California 94043 USA; ⁵Genomics Institute of the Novartis Research Foundation (GNF), San Diego, California 92121 USA.

⁶Corresponding author.

E-MAIL rob_lipshutz@affymetrix.com; FAX (408) 481-0422.

The TAG-SBE genotyping method pairs the extension primer for each marker with a unique tag sequence, allowing the deconvolution of multiplexed preparations on a single high-density probe array (Fig. 1). The TAG-SBE approach can also be multiplexed both at the primary PCR and the SBE steps (see below). The resulting hybridization pattern from a typical TAG-SBE assay is shown in Figure 2A. The intensities of the two fluorophores used are measured and corrected for background and spectral overlap. The quantitative hybridization results are then used to make genotype calls (Figure 2B).

We first tested whether SBE methods for genotyping could be simplified. Previously published SBE methods such as minisequencing (Pastinen et al. 1997,1998; Syvanen 1998) and genetic bit analysis (Nikiforov et al. 1994; Head et al. 1997) required that double-stranded templates be converted to single-stranded templates prior to the base extension reaction [although double-stranded templates have been successfully used in fluorescence energy transfer-based SBE assays (Chen et al. 1997)]. We compared the TAG-SBE results obtained with three SNP markers using

both single-stranded and double-stranded PCR products as templates, and found similar two-color intensity ratios and no significant differences in the absolute hybridization signal intensities. Thus, for all subsequent analyses, and the assays described here, double-stranded PCR templates were used in the SBE reactions.

To test the robustness, accuracy, and efficiency of the TAG-SBE method, we developed genotyping assays for a subset of the 874 SNPs that were identified recently in a large-scale polymorphism screen of 75 hypertension candidate genes (Halushka et al. 1999). Of these, we chose 171 SNPs in 68 genes, focusing on SNPs likely to have a functional significance: We chose SNPs in promoter regions, at splice junctions, and those that altered protein sequence. PCR primers were designed and tested individually for each of the 171 SNP-containing genomic regions. Of these, eight (4.7%) failed to amplify, and SBE primers were designed for the remaining 163 SNPs. We did not attempt to rescue the failed PCRs at this point, but this could be done if needed by reselecting primers or through a modification of the standard PCR conditions. For six of the 163 SNPs, SBE primers were designed for both the forward

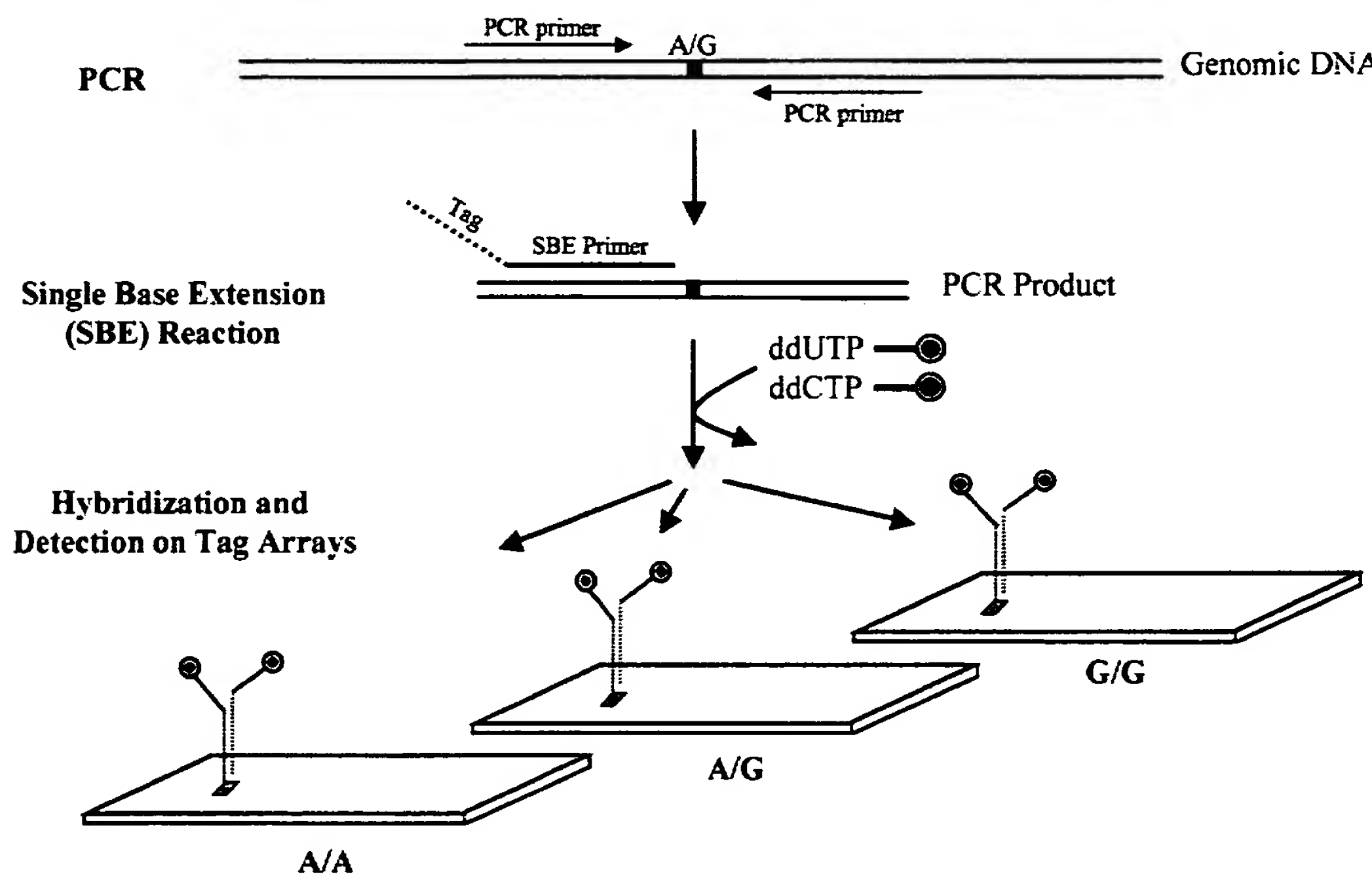


Figure 1 TAG-SBE genotyping assay. Marker-specific primers are designed for amplification of each SNP from genomic DNA (Wang et al. 1998); all SNPs with the same pair of variant bases (e.g., A/G SNPs) are pooled. The double-stranded PCR products serve as templates for the SBE reaction. Each SBE primer is chimeric with a 5' end complementary to a unique tag synthesized on the array and a 3' end complementary to the genomic sequence and terminating one base before a polymorphic SNP site. Thus, each SBE primer is uniquely associated with a specific tag (location) on the array. SBE primers corresponding to multiple markers are added to a single reaction tube and extended in the presence of pairs of ddNTPs labeled with different fluorophores; for example, an A/G bi-allelic marker is extended in the presence of biotin-labeled ddUTP and fluorescein-labeled ddCTP. The labeled multiplex SBE reaction products are pooled and hybridized to the tag array. Three hybridization patterns are shown, corresponding to three genotypes AA, AG, and GG.

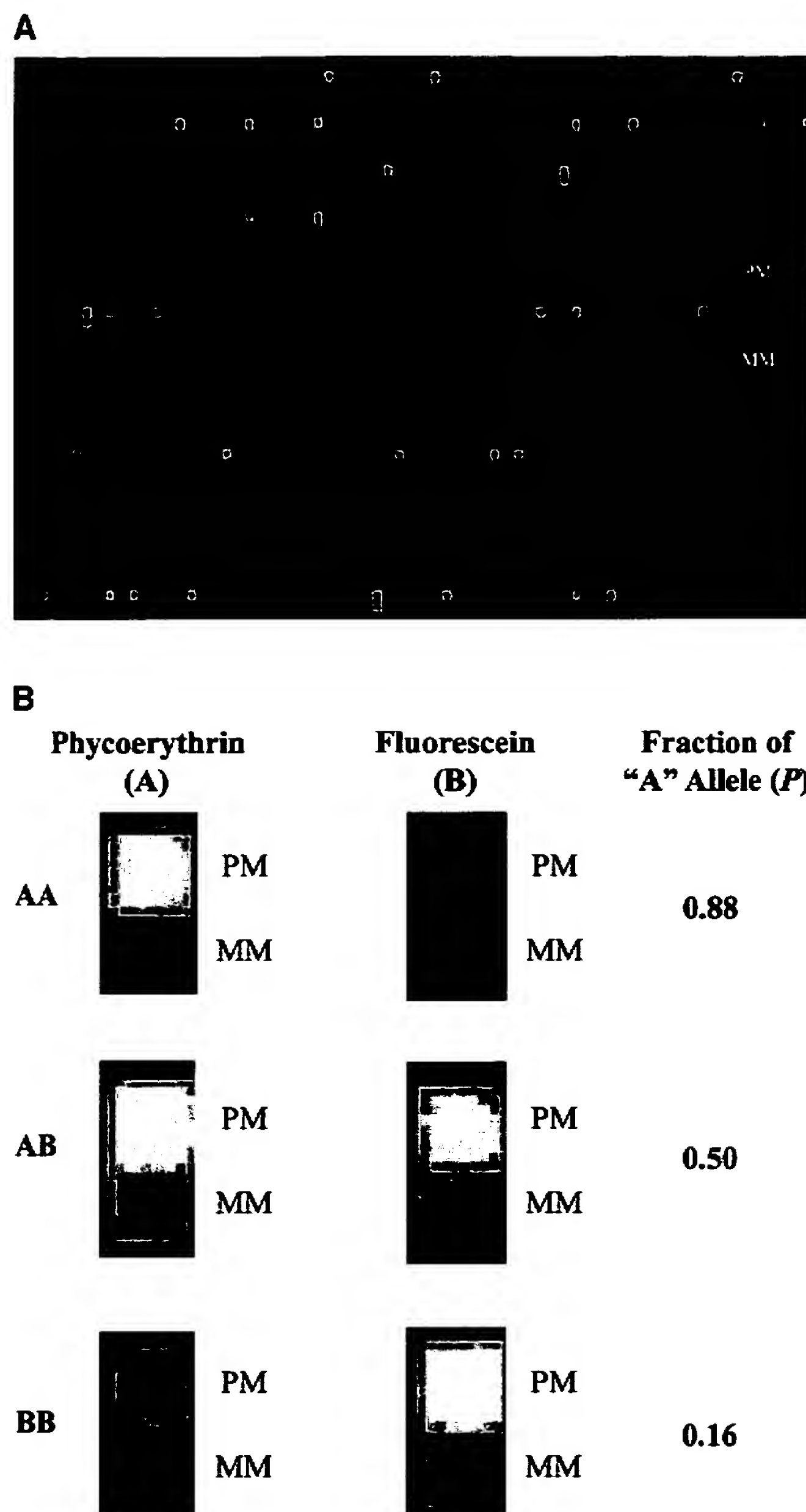


Figure 2 (A) Fluorescence image of a small portion of an oligonucleotide tag array following hybridization of 77 labeled SBE primers. The entire array contains >32,000 20-mer tag probe pairs. The physically adjacent PM and MM probes for a single tag probe pair are labeled. (B) The fluorescence intensity pattern for a tag probe pair showing the presence of an AA homozygote, an AB heterozygote, and a BB homozygote, and the computed relative allele fraction value $P = (PM - MM)_{\text{fluorescein}} / [(PM - MM)_{\text{fluorescein}} + (PM - MM)_{\text{phycoerythrin}}]$. Because of the partial overlap of the emission spectra of fluorescein and phycoerythrin, there is some spillover of fluorescein signal into the phycoerythrin emission channel. Background signals are subtracted and corrections for spectral overlap are applied prior to the quantitative genotyping analysis.

and reverse strands. Nine multiplex PCR and SBE reactions were designed with 9–28 markers in each set. Of the 163 SNP markers tested, 21 SNPs (12.9%) were further eliminated because they consistently produced poor signals in multiple samples tested. These failures were systematic, and were the result of poor amplification in the multiplex PCR or SBE reactions, or poor hybridization behavior on the array. It has been shown previously that roughly one out of 10 tag sequences do not hybridize sufficiently well on arrays of this type (Winzeler et al. 1999). Although these SNPs may be rescued by primer or protocol changes, repooling, using the opposite strand extension primer, or simply linking the primer to a different tag sequence (from which there are many to choose), we have not attempted further optimization of these 29 (8+21) markers. The remaining 142 markers in 62 genes were used in subsequent genotyping experiments. The 142 SNPs used, the genes involved and other details of the polymorphisms, and the designed primers are listed in a table located in the online supplement (note that the first 20 bases of the SBE primers listed in the table are complementary to the tag probes on the array). Additional information on these SNPs can be found in dbSNP (<http://www.ncbi.nlm.nih.gov/SNP/>) or at <http://genome.cwr-u.edu/candidates/snps.html> (Halushka et al. 1999).

To test the reproducibility of the TAG-SBE assay, we performed the multiplex PCR, SBE reactions, and the array hybridization experiments in duplicate for four independent samples. A high correlation between the hybridization signals of the replicate measurements ($R^2 = 0.92$ for fluorescein signals and $R^2 = 0.93$ for phycoerythrin signals) was observed for the 142 SNPs. More importantly, there were no discrepancies in genotyping calls between the duplicate measurements.

We next used tag arrays to obtain the genotypes for all 142 SNPs in 44 unique DNA samples. Hybridization signals sufficiently above background were obtained for 96.5% (6029/6248) of the 6248 (142 × 44) possible calls. Based on the two-color signal intensity ratios, distinct genotype clusters were obtained for ~80% of the markers (Fig. 3). We used a combination of automatic software analysis and blind manual editing to assign genotypes for all 142 markers in the 44

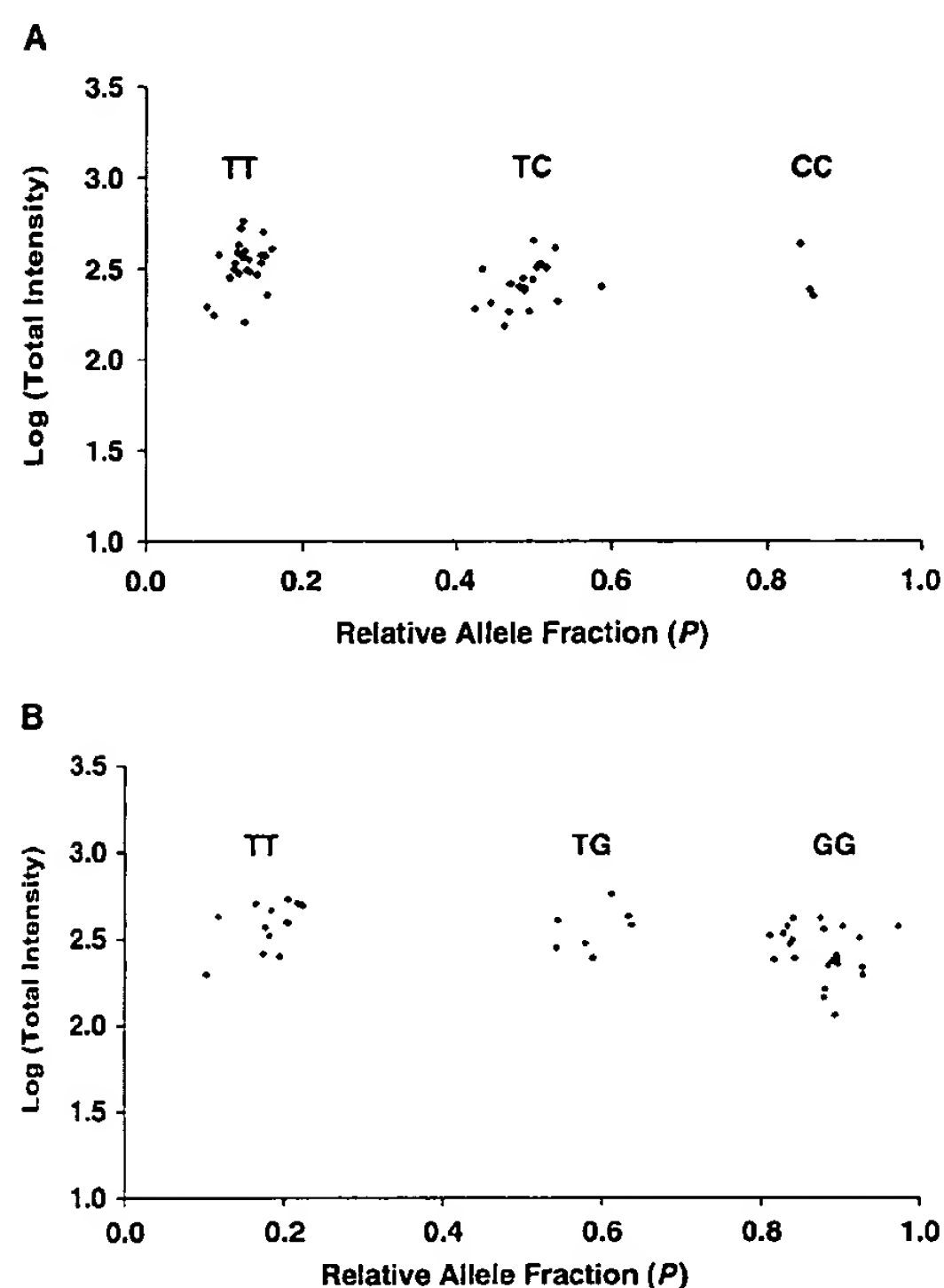


Figure 3 Cluster analysis of tag array hybridization results for 44 individuals at SNP marker (A) ANPex3.33 and (B) SELP.25. The logarithm of total fluorescence intensity $[(PM-MM)_{\text{fluorescein}} + (PM-MM)_{\text{phycoerythrin}}]$ for each of the 44 hybridizations is plotted against the calculated relative allele fraction value P . The three distinct clusters observed correspond to the genotypes T/T, T/C, and C/C for marker ANPex3.33, and T/T, T/G, and G/G for marker SELP.25.

samples. For five of the six SNPs that had both forward and reverse SBE primers, identical genotypes were obtained from both strands in all 44 individuals (i.e., complete concordance in 220 paired tests). For one SNP (DCP1EX13.138), clear hybridization results were obtained for the forward primer, but the results were inconclusive for the reverse SBE primer and therefore calls for that strand were not made (i.e., one strand

yielded clear results while the other produced a “no call”). In no cases did the two strands give contradictory results. This experiment indicates that either strand (or both) can be used for TAG-SBE analysis of the majority of the markers, and that for some markers, one strand may be more informative than the other. As described above, these assays were not fully optimized and we anticipate that it is possible to increase the overall genotyping yield further.

To determine the accuracy of the method, we used gel-based DNA sequencing to determine the genotypes of three individuals (a subset of the 44 persons studied earlier) at 133 loci. Comparison of the 355 paired gel-based and TAG-SBE genotype calls showed a total of 17 discrepancies involving seven different markers (see Table 1), a 4.8% discordance rate. Some of these discrepancies involved cases where one method made a homozygote call while the other method called a heterozygote. But there were also cases in which the gel-based sequencing and array-based genotyping yielded opposite homozygote genotype calls; we suspect systematic mispriming of the SBE primer to adjacent similar sequences as the likely cause of the discrepancy. Designing an SBE assay using primers for the other strand may be sufficient to solve the problem in most cases.

The quantitative nature of the two-color TAG-SBE measurements suggests the possibility of using pooled DNA samples to estimate allele frequencies and screen large numbers of loci for allele frequency differences between groups of phenotypically distinct individuals (Shaw et al. 1998 for microsatellite markers; Syvanen et al. 1993; Hacia et al. 1998 for SNP markers). To test this, we first synthesized two artificial SBE templates and performed controlled mixing experiments. As shown in Figure 4, the intensity ratio of the two fluorophores and the template concentration ratio are highly correlated over a 100-fold concentration range. We further tested the TAG-SBE assay performance with pooled DNA samples. Genomic DNA from five, 10, and 20 individuals with known genotypes was pooled and treated the same way as the DNA samples from individuals in all subsequent PCR amplification, SBE reac-

Table 1. Discrepancies Between Genotyping Calls with Gel-based Sequencing and the Array-based Method

| SNP name | WT allele | Mutant allele | Gel-based sequencing | | | Array-based assay | | | Discrepancies |
|---------------|-----------|---------------|----------------------|-------|--------|-------------------|--------|--------|---------------|
| | | | 904889 | 90896 | 904957 | 904889 | 904896 | 904957 | |
| ACEEX17.19 | C | A | C/C | A/A | A/C | C/C | C/C | C/C | 2 |
| CYP11B2EX6.91 | T | C | T/C | T/C | T/C | T/T | T/T | T/T | 3 |
| CYP11B2BX7.65 | T | C | T/C | C/C | C/C | T/T | T/T | T/T | 3 |
| GLUT4EX3.112 | C | G | G/G | G/G | G/G | C/C | C/C | C/C | 3 |
| GALNREX1.553 | G | C | G/C | G/G | G/C | G/G | G/G | G/G | 2 |
| ICAM1EX6.254 | G | A | G/G | G/G | G/G | A/A | A/A | A/A | 3 |
| GMP-140.25 | T | G | G/G | G/G | G/G | G/G | T/G | G/G | 1 |

tion, and chip hybridization steps. In general, the observed allele frequencies were related directly to the values expected based on the known genotypes of the individuals in the pool (Fig. 5), and relatively small differences in allele frequency could be reliably detected for many markers. This strategy may be used to estimate allele frequencies in populations and to scan large numbers of markers for allele-frequency differences while greatly reducing the number of individual measurements required for association studies designed to detect genetic differences between groups of individuals with phenotypic differences. The minimum detectable allele-frequency differences and the maximum number of markers that can be genotyped in parallel remain to be determined.

Our approach combines the parallelism and flexibility of a standardized high-density oligonucleotide array readout with the enhanced fidelity of enzymatic primer extension reactions. Using a standard array of generic tags eliminates the need to design and manufacture custom arrays for specific sets of markers, as only the PCR and extension primers need to be customized. Furthermore, the tag-based approach uses as few as one or two oligonucleotide probes per marker rather than the 56 probes used previously on variant detector arrays (VDAs) (Wang et al. 1998). The standard tag array could also be used in combination with other genotyping approaches including multiplex oligonucleotide ligation assays (OLA) (Delahunty et al. 1996; Tobe et al. 1996; Chen et al. 1998), invasive cleavage of oligonucleotide probe assays (Lyamichev et al. 1999), and allele-specific PCR methods (Newton et al. 1989; Lo et al. 1991).

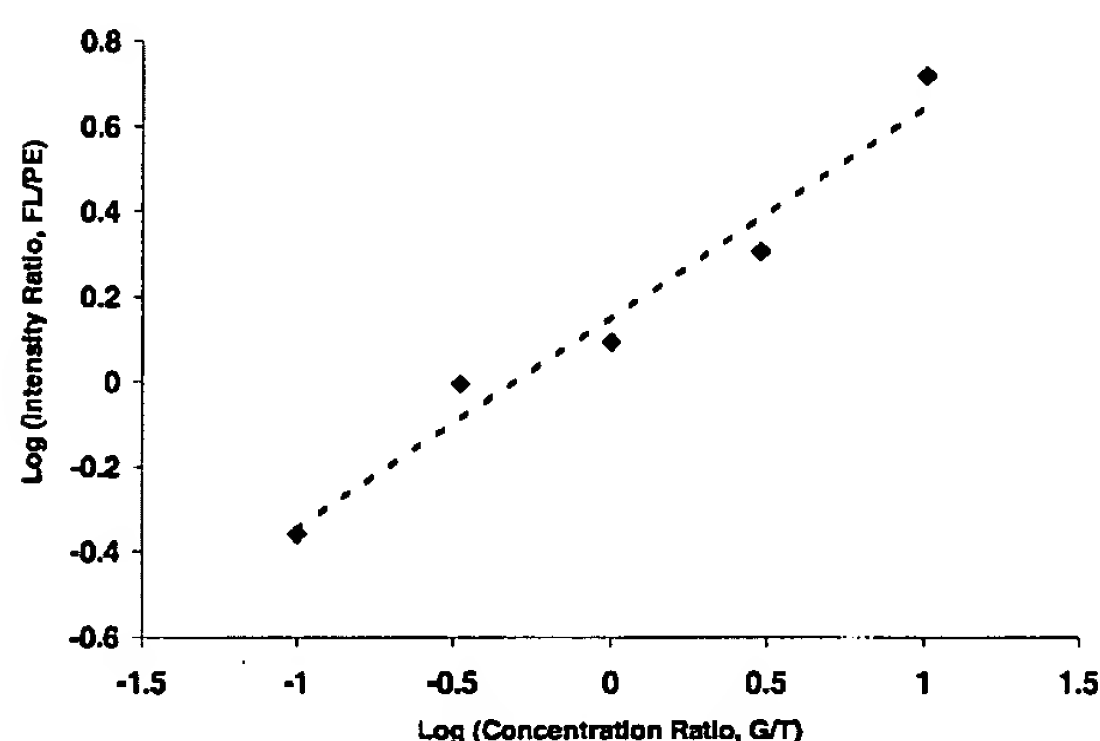


Figure 4 Quantitative allele frequency estimation based on two-color analysis of synthetic mixed templates. The two templates were mixed in the ratios of 1 nM/10 nM, 1 nM/3 nM, 1 nM/1 nM, 3 nM/1 nM, and 10 nM/1 nM, respectively. The logarithm of intensity ratios of the two colors (Y-axis) are plotted against the logarithm of concentration ratios of the two mixed templates (X-axis). FL, fluorescein intensity; PE, phycoerythrin intensity; G/T, concentration ratio of template G to template T.

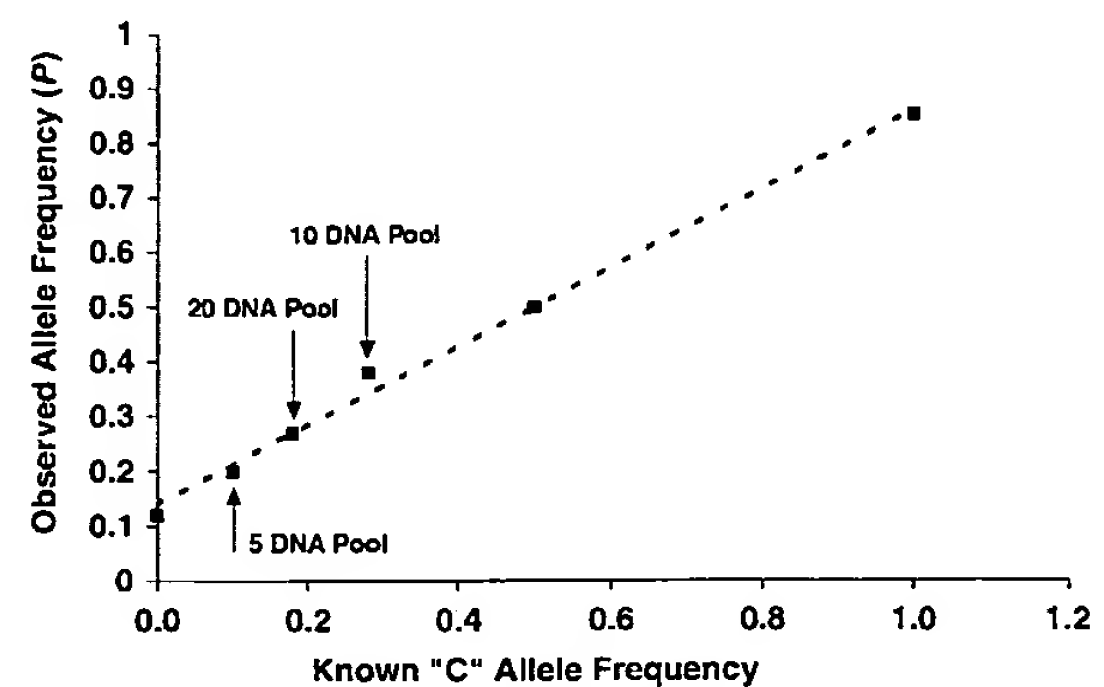


Figure 5 Allele frequency estimation for individual homozygotes, heterozygotes, and collections of multiple individuals at the SNP marker ANPex3.33. For the pooled samples, genomic DNA from a group of 5, 10, and 20 individuals (C-allele frequencies of 0.10, 0.28, and 0.18, respectively) was pooled in equal amounts and treated in the same way as the samples from single individuals. The observed allele fraction value P is plotted against the known C allele frequency, along with the best fit line as a guide to the eye. The line intercepts the Y-axis above the origin, and this systematic offset is the result of a small amount of cross-hybridization and misincorporation of the wrong base in the two-color SBE reaction. A correction can be applied to the data following the observation of pure genotypes to obtain a more accurate estimate of the absolute allele frequencies.

The experiments described here used only a small fraction of the 32,000 tags synthesized on the array and have not taken full advantage of the multiplexing possibilities. Our previous experience with developing highly discriminating sets of oligonucleotide probes for yeast gene expression measurements and genotyping, suggests that it should be possible to use a large fraction of the 32,000 tags on the array in a single experiment (Wodicka et al. 1997; Winzeler et al. 1998). A set of three such arrays would allow the determination of nearly 100,000 genotypes. The current array was synthesized using 30 μm features on an 8 \times 8 mm chip. A single, 12.8 \times 12.8 mm array with 24 μm features could interrogate 128,000 SNPs at a time. Physically smaller arrays with fewer tags may also be useful. Scaling down the array size to 2 \times 2 mm, an array containing 24 μm features could encode over 3000 tags and accommodate many important genotyping applications in which more markers may not be necessary. In addition, multiple sets of tags can be associated with each locus-specific extension primer in separate reactions (pooled for hybridization). In this manner, a single array could be used to analyze the same loci from multiple individuals at once.

The highly parallel nature of oligonucleotide arrays and their ability to interrogate complex mixtures of nucleic acids enables significant flexibility in the design of genotyping assays. Simple calculations suggest that the cost of amplification and labeling reactions can be a significant barrier to the broad use of

large-scale genotyping methods. The multiplex sample preparations demonstrated here permit significant reductions in reagent use. Thus, multiplexing both specific genomic amplifications and SBE reactions reduced the 284 reactions needed for the 142 SNPs to only 18 reactions. This 16-fold reduction can be extended by pooling strategies. The current scheme uses two colors and requires six separate SBE reactions. The use of four colors would allow a single-tube reaction, with associated increases in efficiency and reduction of genotyping costs.

METHODS

Sample Collection and DNA Isolation

DNA samples from 44 individuals were collected as part of the ongoing GenNet network of the National Heart, Lung, and Blood Institute Family Blood Pressure Program. The sampling scheme was designed to ascertain nuclear families through a hypertensive proband. Samples were collected under informed consent and IRB approval at each of two field centers in Tecumseh, MI and Maywood, IL. DNA was extracted from buffy coats isolated from 5 to 10 ml of whole blood using a standard salting-out method and the PureGene kit (Gentra Systems). For the pooling experiments, genomic DNA from five, 10, and 20 individuals was pooled in equal amounts, and treated like single DNA samples in subsequent PCR amplifications, SBE reactions, and chip hybridizations.

Primer Design

For each SNP, primary PCR amplification primers were designed as described previously (Wang et al. 1998). The SBE primers were designed so that the 3' end terminates one base before the polymorphic site. The Primer 3.0 software (<http://www.genome.wi.mit.edu/cgi-bin/primer/primer3.cgi>) was modified and used to pick SBE primers at a predicted length of 20 nucleotides (range: 16–26) and melting temperature of 57°C (range: 53°C–64°C). SBE primers were picked from the forward direction first (i.e., 5' to the SNP), the reverse direction being used when a suitable primer could not be chosen for the forward direction.

Multiplex PCR

Specific amplification of the genomic regions containing the 142 SNPs was achieved with nine multiplex PCR reactions, each containing 50 ng of human genomic DNA, 0.5 μ M of each primer, 1 mM deoxynucleotide triphosphates (dNTPs), 10 mM Tris-HCl (pH 8.3), 50 mM KCl, 5 mM MgCl₂ and 2 units of AmpliTaq Gold (Perkin Elmer) in a total volume of 25 μ l. PCR was performed on a Thermo Cycler (MJ Research) with initial denaturation of the DNA templates and Taq enzyme activation at 96°C for 10 min, followed by 40 cycles of denaturation at 94°C for 30 sec, 57°C for 40 sec, and 72°C for 90 sec. The final extension reaction was at 72°C for 10 min.

SBE Template Preparation

One μ l of Exonuclease I (10 U/ μ l, Amersham Life Science) and 1 μ l of Shrimp Alkaline Phosphatase (1 U/ μ l, Amersham Life Science) were added to 25 μ l PCR products and incubated at 37°C for 1 hr. The enzymes were inactivated at 100°C for 15 min. The enzymatically treated samples were applied to an

S-300 column (Pharmacia) to further remove residual PCR primers and dNTPs. The buffer was replaced with ddH₂O.

Multiplex SBE Reaction

SBE reactions were carried out in 33 μ l reactions using 6 μ l of the template (see above), 1.5 nM of each SBE primer, 2.5 Units of Thermo Sequenase (Amersham), 52 mM Tris-HCl (pH 9.5), 6.5 mM MgCl₂, 25 μ M of fluorescein-N6-d-dNTPs (New England Nuclear), 7.5 μ M biotin-N6-d-dUTP or biotin-N6-d-CTP or 3.75 μ M biotin-N6-d-dATP, and 10 μ M of the other cold ddNTPs. Extension reactions were carried out on a Thermo Cycler (MJ Research) with 1 cycle at 96°C for 3 min, then 45 cycles of 94°C for 20 sec and 58°C for 11 sec. After SBE reactions, the products of the nine reactions from each sample were combined and mixed with 30 μ l of 100 μ g/ml glycogen (Boehringer Mannheim), 18.75 μ l of 8 M LiCl (Sigma), and 1.1 ml of prechilled (–20°C) ethanol (200 proof), and precipitated by centrifugation (Eppendorf centrifuge 5415C) for 15 min at room temperature; precipitated samples were dried at 40°C for 40 min and resuspended in 33 μ l ddH₂O.

Tag Array Design and Hybridization

For each tag sequence, two probes were synthesized on the array: one matches the designed-tag sequence exactly (PM probe) and the other being identical except for a single base difference in the central position (MM probe). The mismatch probe serves as an internal control for hybridization specificity and enables effective subtraction of background and cross-hybridization signals. Over 32,000 20-mer tag probes and their mismatch partners were chosen (Shoemaker et al. 1996) and fabricated on 8 × 8 mm arrays. Each probe (feature) occupies an area of 30 × 30 μ m, which contains ~10⁷ copies of the chosen 20-mer oligonucleotide. Sets of 100 arrays were synthesized together on a single glass wafer.

The labeled SBE reaction products were denatured at 95°C–100°C for 10 min and snap cooled on ice for 2–5 min. The tag array was prehybridized with 6 × SSPE-T [0.9 M NaCl, 60 mM NaH₂PO₄, 6 mM EDTA (pH 7.4), 0.005% Triton X-100] and 0.5 mg/ml BSA for a few minutes, then hybridized with 120 μ l hybridization solution (shown below) at 42°C for 2 hr on a rotisserie (at 40 RPM). The hybridization solution consisted of 3M TMACl (tetramethylammonium chloride), 50 mM MES [2-(N-morpholinoethanesulfonic acid) sodium salt] (pH 6.7), 0.01% of Triton X-100, 0.1 mg/ml of herring sperm DNA, 50 pM of fluorescein-labeled control oligo, 0.5 mg/ml of BSA (Sigma) and 29.4 μ l-labeled SBE products (see above) in a total volume of 120 μ l.

After hybridization, the arrays were rinsed twice with 1 × SSPE-T for ~10 sec at room temperature, then washed with 1 × SSPE-T for 15–20 minutes at 40°C on a rotisserie at 40 RPM. The arrays were washed 10 times with 6 × SSPE-T at 22°C on a fluidics station (FS400, Affymetrix) and then stained at room temperature with 120 μ l staining solution [2.2 μ g/ml streptavidin R-phycoerythrin (Molecular Probes), and 0.5 mg/ml acetylated BSA, in 6 × SSPE-T] and mixed on a rotisserie for 15 min at 40 RPM. After staining, the arrays were washed 10 times with 6 × SSPE-T on the fluidics station at 22°C. The arrays were scanned on a confocal scanner (Affymetrix) and fluorescence at 530 nm (fluorescein), and 560 nm (phycoerythrin) was collected with a spatial resolution of 60–70 pixels per feature. GeneChip software (Affymetrix) was used to convert image files into digitized files for further data analysis.

Genotype Determination

For a given marker (at a given tag probe position), the fluorescence intensity of each of the two fluorophores (fluorescein and phycoerythrin) was corrected for background and nonspecific hybridization by subtracting the intensity at the MM from that of the PM; negative values of PM-MM were treated as zero. Because of the overlap between the emission spectra of the two fluorophores, a fraction of the fluorescein signal (7.6%) was subtracted from the signal seen in the phycoerythrin channel (Hacia et al. 1998). A metric P which estimates the relative amount of each allele in the target mixture was computed as the relative proportion of the corrected intensities [fluorescein/(fluorescein+phycoerythrin)]. To define genotype clusters for each SNP (see Figure 3), the P values associated with each sample were sorted, and ranges corresponding to the three SNP genotypes were computed using an algorithm based on empirical observations across many genotyping experiments. The purpose of this algorithm is to identify well-separated ranges of experimental values that correspond to distinct genotypes. The specific algorithm employed here used the following rules: (1) At most four values (outliers), about 10% of the total data may be excluded from the computed ranges; (2) each pair of ranges must extend over an area of ≥ 0.3 and all three ranges must extend over ≥ 0.5 ; (3) individual ranges must be separated by a gap of ≥ 0.1 ; (4) the width of a single range may be ≤ 0.4 . A "goodness" of fit statistic computed as $1 - (\text{sum of range widths}/\text{total range}) - (\text{number of outliers}/10)$ was maximized for the set of ranges chosen.

Quantitative Allele Analysis

Two templates, template-T (5'-TGCTGAATATTCAGATTCTCTAGTGCTACCTGAAAGATCCTG-3') and template-G (5'-TGCTGAATATTCAGATTCTCGAGTGCTACCTGAAAGATCCTG-3') were synthesized. They were identical except at a single (21st) position: T in template-T, and G in template-G. The two templates were mixed in the ratios of 1 nM/10 nM, 1 nM/3 nM, 1 nM/1 nM, 3 nM/1 nM, and 10 nM/1 nM, respectively. The following five distinct SBE primers, 5'-TGCGATTCTTTGCCGTCAGGCAGGATCTTTCAGGTAGCACT-3', 5'-GGCGAAGTTCCTCTAGTGTTTCAGGATCTTTCAGGTAGCACT-3', 5'-GGCCTCGGTGTTTCAGCATATCAGGATCTTTCAGGTAGCACT-3', 5'-TGGAGATCGTTGCTTG-TACCCAGGATCTTTCAGGTAGCACT-3', 5'-TGCATTGATTAACTGCGCGCAGGATCTTTCAGGTAGCACT-3', were added separately to five SBE reactions containing the five types of mixed templates. The SBE primers were extended in the presence of biotin-labeled ddATP and fluorescein-labeled ddCTP, pooled, and hybridized to a tag array.

Gel-based Automated DNA Sequencing

To independently confirm the genotypes called using the TAG-SBE assay, three samples (904957, 904896, and 904889) were sequenced for 115 SNPs from the table in the online supplement, using conventional gel-based methods. Samples were amplified for all sites with T7- and T3-tagged primers using standard PCR cycling conditions [2.5 μ l of 20 ng/ μ l DNA, 0.375 μ l of 20 μ M primer (X2), 1.5 μ l of 10 \times PCR buffer, 0.9 μ l 25mM MgCl₂, 0.15 μ l 10 mM dNTPs, 0.25 μ l 10 U/ μ l Taq DNA Polymerase (Sigma), in a total volume of 15 μ l with ddH₂O]. Some products were sequenced directly while others required an M13 nesting strategy because of the close proximity of the polymorphic base and primer end. Samples

from the initial amplification were diluted 1:50 with ddH₂O and amplified with M13F-T7 (5'-TGTAACACGACGGCCAGT-TAATACGACTCACTATAGGGAGA-3') and M13R-T3 (5'-AACAGCTATGACCATGAATTAACCCTCACTAAAGGGAGA-3') primers using standard PCR conditions. All PCR products were cleaned with Exonuclease I (Amersham 0.15 μ l of 10 U/ μ l per well) and Shrimp Alkaline Phosphatase (Amersham, 0.30 μ l of 1 U/ μ l per well) in a volume of 10 μ l. Dye terminator sequencing using an M13R primer (AACAGCTATGACCATG) or T7 primer (TAATACGACTCACTATAGGGAGA) on an ABI377 (Perkin Elmer) using Big Dye (Perkin Elmer) was performed to determine the genotype status for each SNP in each of the three individuals. Trace files were read with Edit View 1.0 (Perkin Elmer) software.

ACKNOWLEDGMENTS

We thank Drs. A. Weder and R. Cooper for DNA sample collection, M. Mittmann and D. Shoemaker for tag selection and array design, D. Stern for construction of array scanners used in this study, K. Bentley for DNA sequencing, and K. Gunderson for helpful discussions. This work was supported by a grant from the Advanced Technology Program of the National Institutes of Standards and Technology (70NANB5H1031) to Affymetrix, and research funds from Case Western Reserve University, University Hospitals of Cleveland, the National Heart, Lung, and Blood Institute (HL54466), and the National Institute of Mental Health (MH60007) to A.C. This research is a contribution of GenNet, a network of the National Heart, Lung, and Blood Institute's Family Blood Pressure Program.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

- Cargill, M., D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, N. Patil, C.R. Lane, E.P. Lim, N. Kalayanaraman, J. Nemesh et al. 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22**: 231-238.
- Chakravarti, A. 1999. Population genetics-making sense out of sequence. *Nature Genet.* **21**: 56-60.
- Chee, M., R. Yang, E. Hubbell, A. Berno, X.C. Huang, D. Stern, J. Winkler, D.J. Lockhart, M.S. Morris, and S.P. Fodor. 1996. Accessing genetic information with high-density DNA arrays. *Science* **274**: 610-614.
- Chen, X. and P.Y. Kwok. 1997. Template-directed dye-terminator incorporation (TDI) assay: a homogeneous DNA diagnostic method based on fluorescence resonance energy transfer. *Nucleic Acids Res.* **25**: 347-353.
- Chen, X., B. Zehnbauser, A. Gnirke, and P.Y. Kwok. 1997. Fluorescence energy transfer detection as a homogeneous DNA diagnostic method. *Proc. Natl. Acad. Sci.* **94**: 10756-10761.
- Chen, X., K.J. Livak, and P.Y. Kwok. 1998. A homogeneous, ligase-mediated DNA diagnostic test. *Genome Res.* **8**: 549-556.
- Collins, F.S., M.S. Guyer, and A. Chakravarti. 1997. Variations on a theme: Cataloging human DNA sequence variation. *Science* **278**: 1580-1581.
- Collins, F.S., A. Patrinos, E. Jordan, A. Chakravarti, R. Gesteland, and L. Walters. 1998. New goals for the U.S. Human Genome Project: 1998-2003. *Science* **282**: 682-689.
- Delahunty, C., W. Ankener, Q. Deng, J. Eng, and D.A. Nickerson. 1996. Testing the feasibility of DNA typing for human identification by PCR and an oligonucleotide ligation assay. *Am. J. Hum. Genet.* **58**: 1239-1246.
- Giaever, G., D.D. Shoemaker, T.W. Jones, H. Liang, E.A. Winzeler, A.

- Astromoff, and R.W. Davis. 1999. Genomic profiling of drug sensitivities via induced haploinsufficiency. *Nature Genet.* **21**: 278–283.
- Hacia, J.G., K. Edgemon, B. Sun, D. Stern, S.P. Fodor, and F.S. Collins. 1998. Two color hybridization analysis using high density oligonucleotide arrays and energy transfer dyes. *Nucleic Acids Res.* **26**: 3865–3866.
- Halushka, M., J.-B. Fan, K. Bentley, L. Hsie, N. Shen, A. Weder, R. Cooper, R. Lipshutz, and A. Chakravarti. 1999. Patterns of single nucleotide polymorphisms in candidate genes regulating blood pressure homeostasis. *Nature Genet.* **22**: 239–247.
- Head, S.R., Y.H. Rogers, K. Parikh, G. Lan, S. Anderson, P. Golet, and M.T. Boyce-Jacino. 1997. Nested genetic bit analysis (N-GBA) for mutation detection in the p53 tumor suppressor gene. *Nucleic Acids Res.* **25**: 5065–5071.
- Kuppuswamy, M.N., J.W. Hoffmann, C.K. Kasper, S.G. Spitzer, S.L. Groce, and S.P. Bajaj. 1991. Single nucleotide primer extension to detect genetic diseases: experimental application to hemophilia B (factor IX) and cystic fibrosis genes. *Proc. Natl. Acad. Sci.* **88**: 1143–1147.
- Lipshutz, R.J., S.P. Fodor, T.R. Gingeras, and D.J. Lockhart. 1999. High density synthetic oligonucleotide arrays. *Nature Genet.* **21**: 20–24.
- Lo, Y.M., P. Patel, C.R. Newton, A.F. Markham, K.A. Fleming, and J.S. Wainscoat. 1991. Direct haplotype determination by double ARMS: Specificity, sensitivity and genetic applications. *Nucleic Acids Res.* **19**: 3561–3567.
- Lyamichev, V., A.L. Mast, J.G. Hall, J.R. Prudent, M.W. Kaiser, T. Takova, R.W. Kwiatkowski, T.J. Sander, M. de Arruda, D.A. Arco et al. 1999. Polymorphism identification and quantitative detection of genomic DNA by invasive cleavage of oligonucleotide probes. *Nature Biotech.* **17**: 292–296.
- Newton, C.R., A. Graham, L.E. Heptinstall, S.J. Powell, C. Summers, N. Kalsheker, J.C. Smith, and A.F. Markham. 1989. Analysis of any point mutation in DNA: The amplification refractory mutation system (ARMS). *Nucleic Acids Res.* **17**: 2503–2516.
- Nikiforov, T.T., R.B. Rendle, P. Golet, Y.H. Rogers, M.L. Kotewicz, S. Anderson, G.L. Trainor, and M.R. Knapp. 1994. Genetic bit analysis: a solid phase method for typing single nucleotide polymorphisms. *Nucleic Acids Res.* **22**: 4167–4175.
- Pastinen, T., A. Kurg, A. Metspalu, L. Peltonen, and A.C. Syvanen. 1997. Minisequencing: A specific tool for DNA analysis and diagnostics on oligonucleotide arrays. *Genome Res.* **7**: 606–614.
- Pastinen, T., M. Perola, P. Niini, J. Terwilliger, V. Salomaa, E. Vartiainen, L. Peltonen, and A.C. Syvanen. 1998. Array-based multiplex analysis of candidate genes reveals two independent and additive genetic risk factors for myocardial infarction in the Finnish population. *Hum. Mol. Genet.* **7**: 1453–1462.
- Risch, N. and K. Merikangas. 1996. The future of genetic studies of complex human diseases. *Science* **273**: 1516–1517.
- Shaw, S.H., M.M. Carrasquillo, C. Kashuk, E.G. Puffenberger, and A. Chakravarti. 1998. Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. *Genome Res.* **8**: 111–123.
- Shoemaker, D.D., D.A. Lashkari, D. Morris, M. Mittmann, and R.W. Davis. 1996. Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nature Genet.* **14**: 450–456.
- Syvanen, A.C. 1998. Solid-phase minisequencing as a tool to detect DNA polymorphism. *Methods Mol. Biol.* **98**: 291–298.
- Syvanen, A.C., K. Aalto-Setälä, L. Harju, K. Kontula, and H.A. Soderlund. 1990. Primer-guided nucleotide incorporation assay in the genotyping of apolipoprotein E. *Genomics* **8**: 684–692.
- Syvanen, A.C., A. Sajantila, and M. Lukka. 1993. Identification of individuals by analysis of biallelic DNA markers, using PCR and solid-phase minisequencing. *Am. J. Hum. Genet.* **52**: 46–59.
- Tobe, V.O., S.L. Taylor, and D.A. Nickerson. 1996. Single-well genotyping of diallelic sequence variations by a two-color ELISA-based oligonucleotide ligation assay. *Nucleic Acids Res.* **24**: 3728–3732.
- Wang, D.G., J.-B. Fan, C.J. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, N. Perkins, E. Winchester, J. Spencer et al. 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**: 1077–1082.
- Winzeler, E.A., D.R. Richards, A.R. Conway, A.L. Goldstein, S. Kalman, M.J. McCullough, J.H. McCusker, D.A. Stevens, L. Wodicka, D.J. Lockhart, and R.W. Davis. 1998. Direct allelic variation scanning of the yeast genome. *Science* **281**: 1194–1197.
- Winzeler, E.A., D.D. Shoemaker, A. Astromoff, H. Liang, K. Anderson, B. Andre, R. Bangham, R. Benito, J.D. Boeke, H. Bussey et al. 1999. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**: 901–906.
- Wodicka, L., H. Dong, M. Mittmann, M.H. Ho, and D.J. Lockhart. 1997. Genome-wide expression monitoring in *Saccharomyces Cerevisiae*. *Nat. Biotechnology*. **15**: 1359–1367.

Received January 5, 2000; accepted in revised form March 29, 2000.

Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy

Daniel D. Shoemaker¹, Deval A. Lashkari¹, Don Morris², Mike Mittmann² & Ronald W. Davis¹

A quantitative and highly parallel method for analysing deletion mutants has been developed to aid in determining the biological function of thousands of newly identified open reading frames (ORFs) in *Saccharomyces cerevisiae*. This approach uses a PCR targeting strategy to generate large numbers of deletion strains. Each deletion strain is labelled with a unique 20-base tag sequence that can be detected by hybridization to a high-density oligonucleotide array. The tags serve as unique identifiers (molecular bar codes) that allow analysis of large numbers of deletion strains simultaneously through selective growth conditions. Hybridization experiments show that the arrays are specific, sensitive and quantitative. A pilot study with 11 known yeast genes suggests that the method can be extended to include all of the ORFs in the yeast genome, allowing whole genome analysis with a single selective growth condition and a single hybridization.

The recent release of the complete nucleotide sequence of *Saccharomyces cerevisiae* provides researchers with the identities of all the genes in this basic eukaryotic cell. However, functional information is available for less than half of the estimated 6,000 genes¹. Identifying the biological function of the thousands of genes discovered by the yeast (and other) genome sequencing projects is one of the main challenges facing researchers as we enter the post-genome era. Determining when a gene is expressed, where the gene product is localized in the cell, and what happens when the gene is disrupted are a few basic experimental questions that need to be addressed for the large numbers of newly discovered genes. There is a clear need for new genome-wide approaches that can provide answers to these questions in a rapid and cost effective manner.

Gene disruption is a powerful tool for determining the biological function of proteins encoded by uncharacterized ORFs. Given the sequence of a yeast ORF, it is possible to generate a precise null mutation (a deletion from start codon to stop codon)^{2,3}. Information about the biological function can be inferred by monitoring the fitness of the resulting deletion strain under a variety of selective growth conditions (for example, exposure to UV light). This approach has been extremely effective for investigating the function of yeast ORFs. Unfortunately, the time and labour involved in analysing individual deletion strains will make it difficult to apply this approach on a genome-wide level.

Strategies have been reported in which gene disruptions⁴ and selections are performed *en masse* using randomly integrated transposons⁵. These methods offer a dramatic increase in the rate of analysing null mutations

compared with the traditional single gene approach. In addition, these novel approaches provide valuable information concerning gene expression⁴, gene product localization⁴ and the functional boundaries of ORFs⁵. However, using these techniques, thousands of ORFs must be individually examined following each selective growth condition to determine which mutants have displayed modified growth characteristics. This labour-intensive analysis step will make it difficult to test a large number of selective conditions on a genome-wide level.

Some of these difficulties have been overcome by a method in which transposons containing unique tags are used to generate insertional mutants⁶. The distinguishing tags allow collections of mutants to be pooled and analysed in parallel through selective growth assays. This approach has the advantage that all three steps (gene disruption, selection and analysis) are performed in parallel. Unfortunately, only 96 strains can be analysed in a single pool due to limitations of the filter-based hybridization assay⁶. In addition, the random nature of the transposon-based mutagenesis requires cloning and DNA sequencing to identify ORFs that are disrupted in strains with interesting phenotypes.

Here, we describe two novel approaches termed "quantitative phenotypic analysis" and "molecular bar-coding," in which directed gene replacement is used to generate individual deletion strains that are each labelled with a distinguishing molecular tag (unique 20 base-pair sequence). Tagged strains are pooled and analysed in parallel through selective growth conditions. The level at which each strain survives a given competitive growth condition can be determined by hybridizing the tags to a high-density oligonucleotide array. This approach has

¹Department of Biochemistry, Beckman Center, Stanford University Medical Center, Stanford, CA 94305, USA

²Affymetrix, 3380 Central Expressway, Santa Clara, California 95051, USA

Correspondence should be addressed to R.W.D.

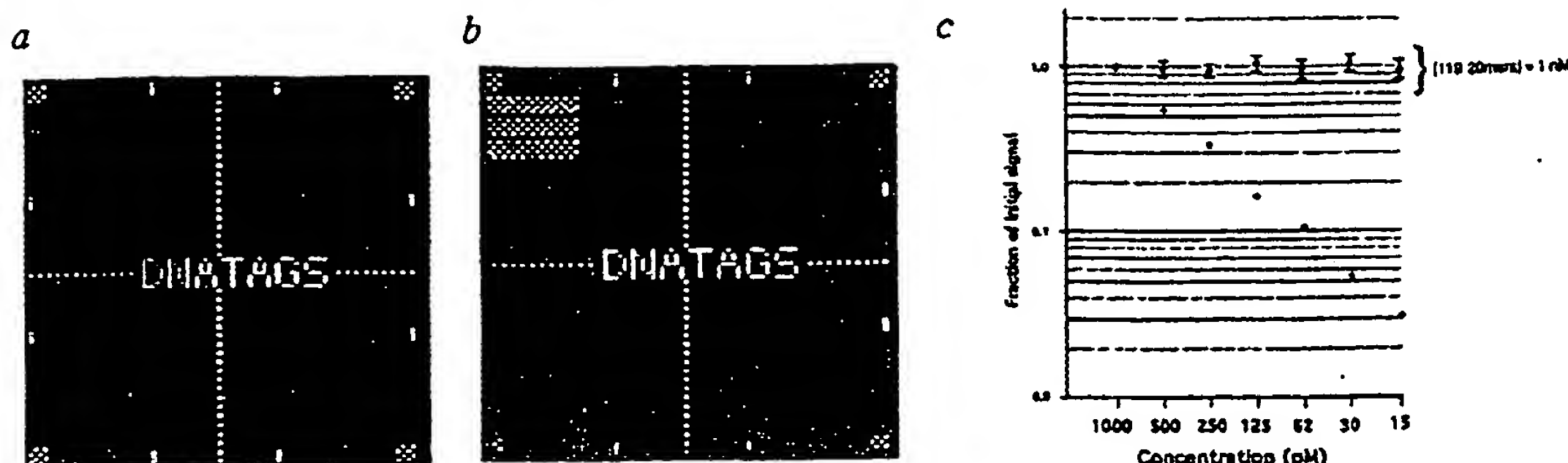
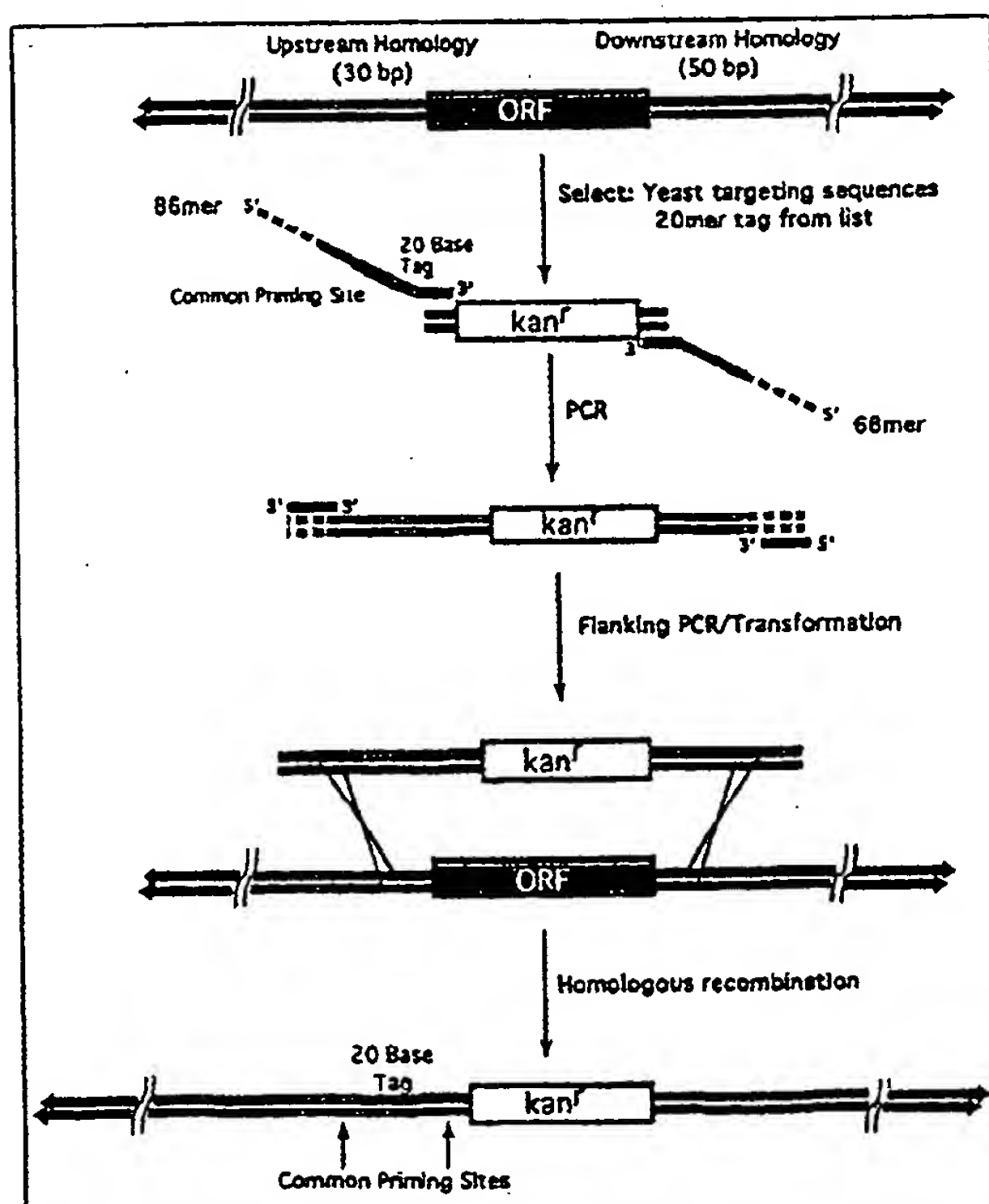


Fig. 1 Scanned images of a high-density oligonucleotide array containing 20mers that are complementary to 4,500 tag sequences. **a**, Hybridization pattern obtained when the array is hybridized with the fluorescently labelled 20mer that is complementary to the control oligonucleotide. Dark areas correspond to the synthesis sites for the 4,500 molecular tags. **b**, Hybridization pattern obtained when the array was hybridized with 120 different fluorescently labelled 20mers along with the control oligonucleotide. **c**, Plot showing the normalized data from a titration study using the pool of 120 fluorescently labelled oligonucleotides. Seven hybridization experiments were performed in which the concentration of one oligonucleotide was varied (1 nM–15 pM) while the remaining 119 oligonucleotides were held constant at 1 nM.

four main advantages: i) large pools of tagged deletion strains (potentially >6,000) can be simultaneously analysed because of the increased sensitivity afforded by fluorescent detection on high-density arrays; ii) deletion strains have ORFs completely deleted, avoiding residual or altered functions associated with truncated products; iii) deletion strains with interesting phenotypes are directly identified without need for further cloning or sequencing; and iv) the quantitative nature of the chip hybridization assay should make it possible to reveal subtle phenotypes that may otherwise be missed.

The steps of molecular bar-coding are as follows:



i) individual deletion strains are generated by replacing specific ORFs with a selectable marker that is linked to a 20 base-pair molecular tag that serves as a unique identifier; ii) large numbers of tagged deletion strains are pooled and grown competitively under various selective conditions (for example, minimal medium, exposure to X-rays, high salt, and so on); iii) the molecular tags are amplified from the surviving strains using a common set of primers and hybridized to a high-density array containing, at defined positions, known oligonucleotides that are complementary to the tag sequences; and iv) the intensities of the hybridization signals for the tags on the array are measured and used to determine the relative abundance of the corresponding deletion strains in the pool. Measuring the relative abundance of the tags at different times during selective growth allows the fitness of each deletion strain in the pool to be quantitatively determined.

Tag design

The molecular tags are 20 base-pair DNA sequences specifically designed to serve as unique identifiers. Tag sequences are as different as possible yet still retain similar hybridization properties to facilitate simultaneous analysis on high-density oligonucleotide arrays. We used an algorithm to select a set of 9,105 maximally distinguished 20mer tag sequences that are predicted to have similar melting temperatures ($61 \pm 5^\circ\text{C}$), no secondary structure and no extensive similarity between any two sequences in the list (>5 mismatches). This set of tags is sufficient for the entire yeast genome.

Fig. 2 PCR-targeting strategy used to generate tagged deletion strains. A dominant selectable marker (*kan^r*) was amplified using a pair of long primers that contained yeast sequences on the 5' end and homology to the marker on the 3' end. One of the oligonucleotides is a 68mer that contains 50 bases of yeast homology and 18 bases of homology to the marker. The other oligonucleotide is an 86mer that contains a 20 base tag and an 18 base tag priming site in addition to the 30 bases of yeast homology and the 18 bases of marker homology. The dashed lines represent missing sequences on the 5' ends of the long unpurified oligonucleotides (see Methods). A second round of PCR is performed with 20mers that are homologous to the ends of the initial PCR product to increase the amount of full length product. The product from the second round PCR is transformed into a haploid yeast strain and homologous recombination results in the replacement of the targeted ORF with the marker, 20mer tag, and tag priming site.

BEST AVAILABLE COPY

High-density oligonucleotide arrays

Hybridization to a high-density oligonucleotide array can detect the presence of a fluorescently labelled tag. We used a combination of photolithography and oligonucleotide synthesis chemistry to generate arrays containing thousands of oligonucleotides complementary to the different 20mer tag sequences^{7,8}. Each element on the array measures 130 × 130 µm and contains approximately 10⁵ oligonucleotides of a given sequence⁹. The location and identity of each sequence on the two-dimensional glass surface is known. Fluorescently labelled DNA is applied to the array and scanning confocal microscopy detects specific hybridization events. A single experiment (30 minute hybridization + 15 minute scan) generates quantitative hybridization information for thousands of different sequences.

To determine the optimal length for the tag sequences, we synthesized and tested high-density arrays containing thousands of specifically designed 8mers, 15mers, and 20mers (1,025, 8,000, and 4,500 oligonucleotides, respectively). The specificity and intensity of the hybridization signals obtained with 20mers were consistently superior to those obtained with the shorter oligonucleotides (data not shown).

For this study, we synthesized high-density arrays containing 20mers complementary to 4,500 of the specially designed tag sequences and synthesized control sequences in the corners, in a cross-hair pattern across the array and in a 'DNA TAGS' configuration. Fig. 1a shows a scanned image of an array hybridized with a fluorescently labelled 20mer complementary to the control sequences on the array. The dark areas on the image correspond to the synthesis sites for the 4,500 tags that did not cross-hybridize with the control oligonucleotide.

To test the hybridization properties of the tag sequences, we synthesized 120 fluorescently labelled oligonucleotides that are complementary to specific tag sequences on the array. These fluorescently labelled 20mers should generate a 10 × 12 checker-board hybridization pattern in the upper left-hand corner of the array. The predicted hybridization pattern was obtained with virtually no cross-hybridization outside the 120 targeted array elements (Fig. 1b). In this hybridization experiment, the signal intensity for some of the tags is higher than others (Fig. 1b). These highly reproducible differences reflect either the sequence dependence of the hybridization reaction and/or variation in the quality of the 120 synthetic, labelled oligonucleotides.

To determine whether the hybridization signals on the array accurately reflect the concentrations of tags in a population, we performed a titration experiment using the pool of 120 labelled oligonucleotides. The concentration of one oligonucleotide was varied in a series of seven hybridization experiments (1 nM, 0.5 nM, 0.25 nM, 0.12 nM, 60 pM, 30 pM, and 15 pM), while the concentrations of the remaining 119 oligonucleotides were held constant at 1 nM. A two-fold change in con-

Table 1 Transformation results for auxotrophic ORFs

| ORF name | ORF size | Targeting efficiency |
|---------------|----------|----------------------|
| ADE1 (M61209) | 305aa | 100% (8/8) |
| ADE2 (M59824) | 571aa | 50% (4/8) |
| ADE3 (M24737) | 945aa | 50% (4/8) |
| ADE4 (M74309) | 509aa | 100% (8/8) |
| ADE5 (X04337) | 802aa | 100% (8/8) |
| ARO2 (X60190) | 378aa | 75% (6/8) |
| ARO7 (M24517) | 256aa | 100% (8/8) |
| TRP2 (K01388) | 507aa | 100% (8/8) |
| TRP3 (K01386) | 484aa | 88% (7/8) |
| TRP4 (X04273) | 380aa | 75% (6/8) |
| TRP5 (V01342) | 707aa | 88% (7/8) |

The ORF name (GenBank accession number), ORF size and targeting efficiency (correct integrations versus total) are shown for each of the 11 auxotrophic ORFs.

centration was detected in the presence of this complex hybridization mixture (Fig. 1c). In addition, the hybridization intensities for the titrated oligonucleotide decreased in a linear fashion over the 70-fold concentration range tested (1 nM to 15 pM).

Generating tagged deletion strains

We used a variation of a previously described PCR targeting strategy^{3,10} to generate the tagged deletion strains (Fig. 2). Normally, a selectable marker (such as *kan*^r) is amplified using a pair of long primers that contain yeast targeting sequences on the 5' end and homology to the marker

on the 3' end. In our modified scheme, a 20-base tag and an 18 base tag priming site are incorporated into one of the oligonucleotides. The PCR-amplified selectable marker is transformed into a haploid yeast strain and homologous recombination results in the replacement of the targeted ORF (start codon to stop codon) with the marker, 20mer tag, and tag priming site. The integrated tag serves as a unique, permanent identifier for that deletion strain.

Pilot study

Gene disruption. To test the PCR targeting strategy (described in Fig. 2), we replaced 11 auxotrophic genes with deletion cassettes containing different pre-selected molecular tags. For each of the transformations, we analysed several G418 resistant colonies using a PCR assay that differentiates between correct and incorrect integration events¹¹. Table 1 shows the transformation results for the 11 auxotrophic genes. The targeting efficiency ranged from 50–100% and there was no correlation between the correct integration frequency and the size of the targeted ORF.

Competitive growth. We tested the feasibility of using the molecular tags to monitor a population of cells using the auxotrophic deletion strains described in Table 1. A deletion pool was generated by combining equal numbers of cells from each of the 11 tagged deletion strains. Genomic DNA extracted from yeast cells in the deletion pool served as the template for a tag amplification reaction. This amplification step serves to increase the concentration of the tags and reduce the complexity of the sample hybridized to the array (mixture of amplified tags versus total yeast genomic DNA). The amplified tags were hybridized to the high-density array which was then washed and scanned. The signals for the 11 tags were intense and highly specific (Fig. 3). Only one out of the 131 tags tested so far has displayed a significant level of cross-hybridization (Fig. 3, yellow asterisk).

The 11 auxotrophic deletion strains provide a good system for testing the bar-coding strategy because the phenotypes are known. In a competitive growth experiment, the relative abundance of these strains should be maintained when the deletion pool is cultivated in complete medium because the strains are predicted to have

BEST AVAILABLE COPY

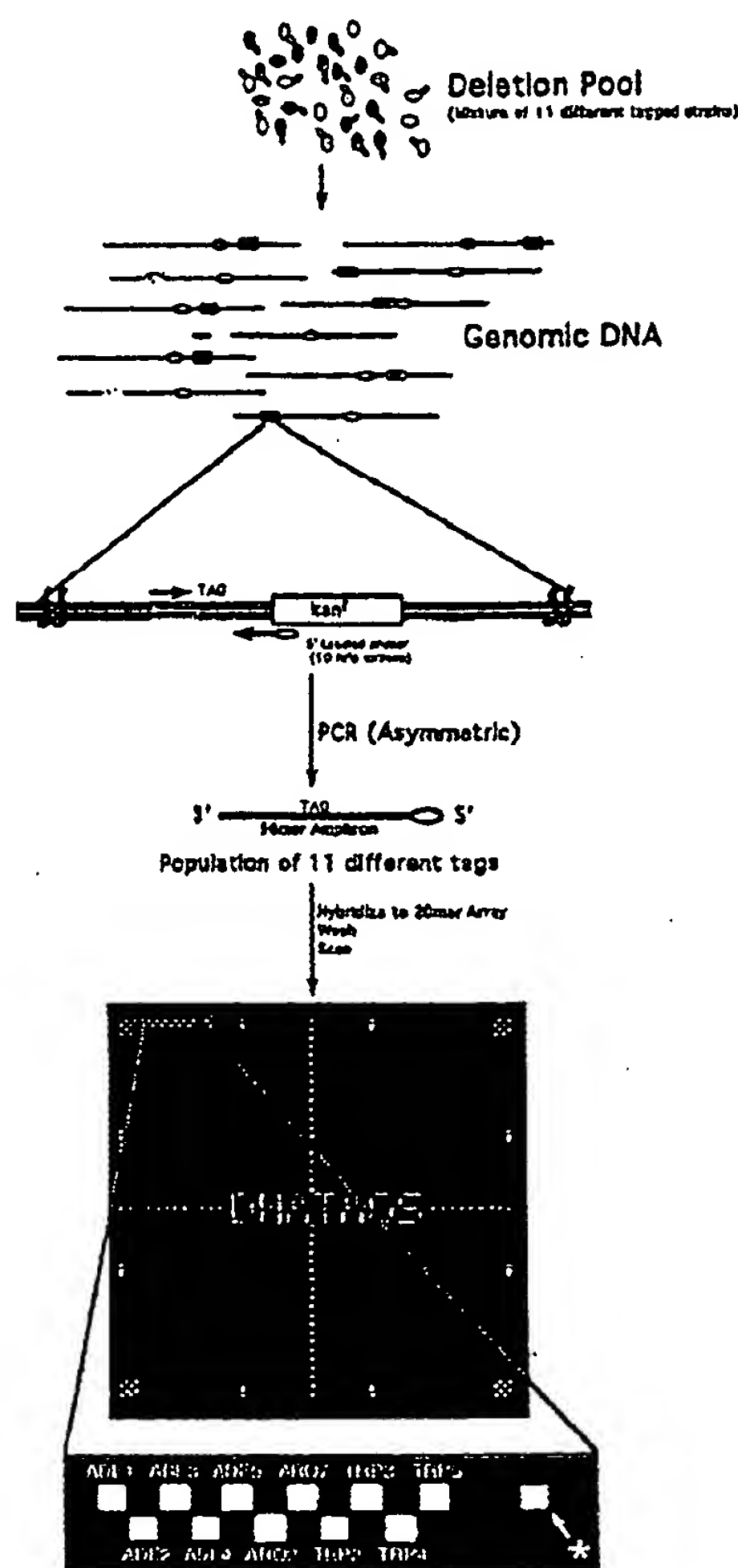


Fig. 3 Tag amplification strategy. Genomic DNA isolated from a deletion pool containing 11 tagged deletion strains was used as template for the tag amplification reaction. The tags were amplified using a pair of primers that were complementary to the common priming sites. The asymmetric PCR generates a mixture of single-stranded fluorescently labelled 56mer tag amplicons that were directly hybridized to the high-density array. A scanned image of the array is shown. A close-up view of the left hand corner shows the exact location of the 11 tags on the array. The asterisk indicates a cross-hybridizing sequence (see text).

similar growth rates. On the other hand, the *ade1Δ-ade5Δ* deletion strains or the *trp2Δ-trp5Δ* and *aro2Δ* deletion strains should be depleted from the population when the pool is grown in media lacking adenine or tryptophan, respectively. To test this prediction, we cultivated representative aliquots of the deletion pool in complete synthetic medium (SDC), medium lacking adenine (SDC-ADE), or medium lacking tryptophan (SDC-TRP). At various time points after inoculation, tags were isolated from the pool and analysed on the array to determine the relative abundance of the different deletion strains in the pool. Similar hybridization

patterns were obtained at each of the time points when tags were isolated from the deletion pool grown in complete medium (Fig. 4a). In contrast, the signal intensities of the tags associated with the *ade1Δ*, *ade2Δ*, *ade3Δ*, *ade4Δ*, and *ade5Δ* deletion strains became progressively weaker when the pool was grown in medium lacking adenine (Fig. 4b). The expected depletion of the five Trp⁻ deletion strains (*trp2Δ*, *trp3Δ*, *trp4Δ*, *trp5Δ* and *aro2Δ*) occurred when the pool was grown in medium lacking tryptophan (Fig. 4c). Figure 4d shows a quantitative representation of the hybridization data from the growth study in medium missing adenine. We measured the signal intensities of the *ade1Δ* and *trp5Δ* deletion strains at various time points during the growth study and plotted the normalized hybridization data (Fig. 4d). The graph clearly shows that the relative abundance of the *trp5Δ* deletion strain was maintained through the competitive growth while the *ade1Δ* deletion strain was depleted from the population (see Methods; data analysis).

Colony colour assay

We used an independent approach to determine whether the signal intensities of the tags on the array accurately reflect the relative abundance of different deletion strains in the pool. The *ade1Δ* and *ade2Δ* deletion strains form red colonies when plated on medium containing limiting amounts of adenine while the other nine deletion strains form white colonies on this medium. Determining the percentage of red colonies provides a direct measurement of the relative abundance of the *ade1Δ* and *ade2Δ* deletion strains in the pool. We cultivated representative aliquots of the deletion pool in complete medium and medium lacking adenine, and at each of the indicated time points, plated several thousand cells from the deletion pool. As predicted, 18% of the colonies taken from the starting population (2/11 deletion strains) were red. This percentage of red colonies remained constant when the deletion pool was grown in complete medium (SDC) for 10 population doublings. In contrast, the percentage of red colonies decreased when the deletion pool was grown in medium lacking adenine (Fig. 4e). The similarity between the plots shown in Figs 4d and 4e verifies that the hybridization signals on the array reflect the actual abundance of the tagged deletion strains in the pool.

Discussion

Determining the biological function for many of the newly identified yeast ORFs will require that thousands of different deletion strains be tested under a large variety of selection conditions¹. Molecular bar-coding is ideal for this task because it allows large numbers of tagged deletion strains to be analysed simultaneously in a highly quantitative fashion. The labour-intensive (genome-wide) step of making the thousands of deletion strains will only have to be performed once, after which all subsequent selections and analyses could be performed in parallel. Furthermore, the quantitative nature of this approach may make it possible to observe subtle growth differences that might otherwise be overlooked. For example, the relative abundance of a single deletion strain with a 5% growth defect (doubling time of 147 minutes versus 140 minutes), grown in a competitive environment with 5,999 other deletion strains

BEST AVAILABLE COPY

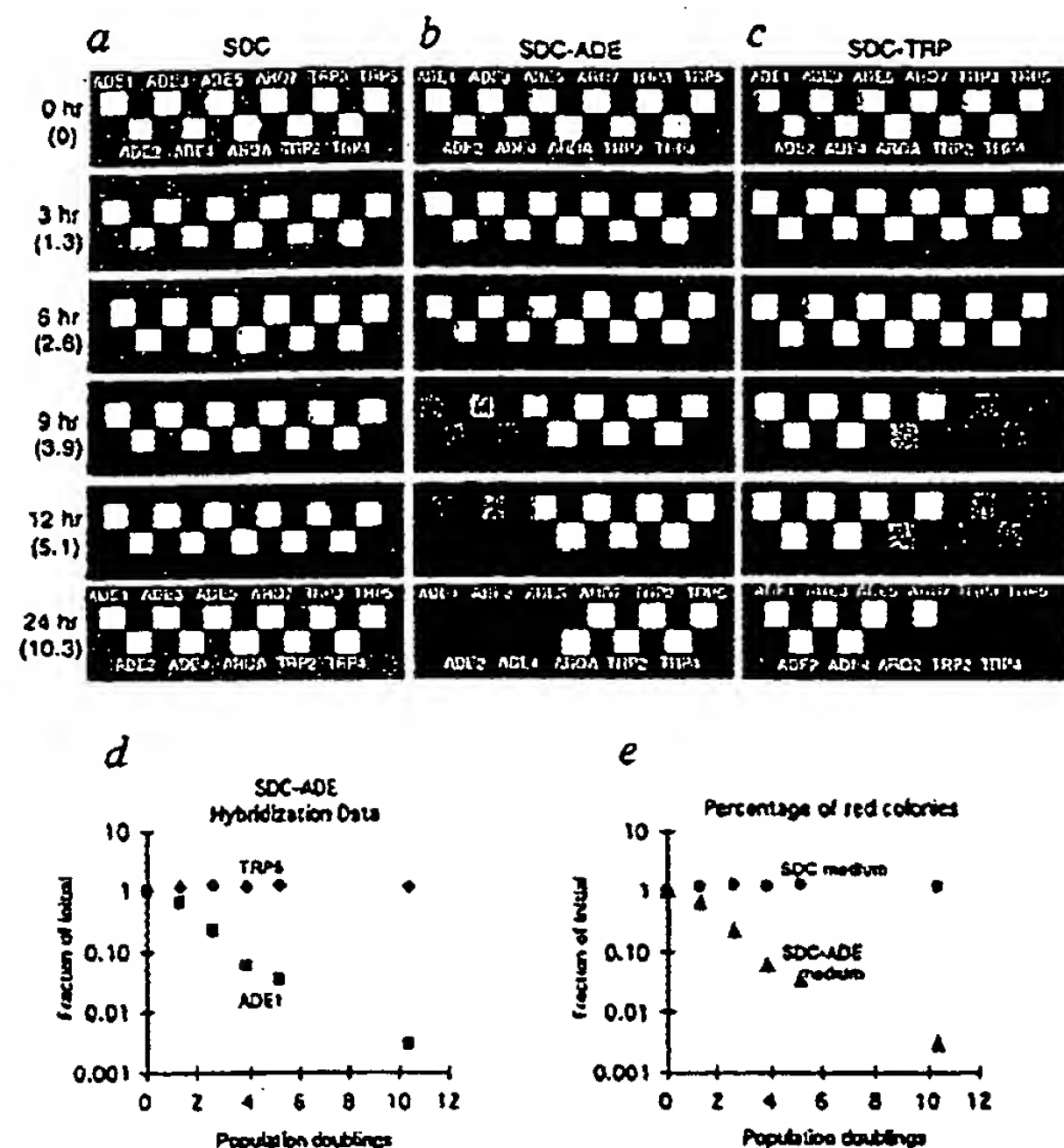


Fig. 4 Competitive growth study using a pool containing 11 tagged auxotrophic deletion strains. **a**, A representative aliquot of the deletion pool was grown in complete synthetic medium (SDC) and tags were isolated at the indicated times (population doublings are shown in parentheses) and analysed on the array. The upper-left hand corner of the array is shown for each of the time points **b**, A similar growth study performed in medium lacking adenine (SDC-ADE) and **c**, medium lacking tryptophan (SDC-TRP). **d**, Plot showing the normalized values for the *ade1Δ* and *trp5Δ* deletion strains from the growth study in medium lacking adenine. **e**, A plot showing the data from the colony colour assay. Aliquots of the deletion pool were grown in complete medium (SDC) and medium missing adenine (SDC-ADE). The relative abundance the *ade1Δ* and *ade2Δ* deletion strains in the pool was determined at each of the indicated time points by measuring the percentage of red colonies. The percentages from each time point were normalized to the values observed in the starting pool.

minimized by using a common pair of primers and by the fact that all of the tags are the same length, have similar base composition and are free of secondary structure. The existing light-directed synthesis technology can generate high-density arrays containing over 65,000 different oligonucleotides¹³, which is well beyond the requirements of this project. Furthermore, the current fluorescent detection system can readily measure concentrations below 1 pM, which should be sufficient for monitoring pools containing more than 6,000 deletion strains (see Methods; pool size). Finally, the initial hybridization results with the 131 tags suggest that cross-hybridization is minimal and largely avoidable. All of the tags will be tested for cross-hybridization before use in the deletion strategy.

The goal of this project is to provide experimental data for thousands of deletion strains in a rapid and cost effective manner. These data, along with the collection of individual deletion strains, will be a valuable resource for investigating the function of genes in yeast.

Additional applications of the molecular bar-coding strategy include the generation of a complete set of tagged deletions in a 'true' wild-type yeast strain, one that has been recently isolated from a known ecological niche. The collection of deletion strains could be released back into their natural environment and the population monitored over time by surveying the PCR-amplifiable tags. This type of analysis would be useful for determining the activity of genes as a function of the environment. Molecular bar-coding may become an invaluable tool in this emerging field of 'molecular ecology'.

The molecular tags will also facilitate the task of keeping track of the thousands of deletion strains. Normally, different deletion strains are identified by labels (or bar-codes) that are placed on the outside of the tubes or micro-titer plates. Incorrect labelling of tubes or putting the wrong strain in the wrong tube are serious problems that can be minimized by having a unique identifier in the actual deletion strains themselves. The identity of any strain could be confirmed by performing a simple tag-specific PCR or by analysing the tag on an array. Furthermore, the fact that the tags are physically linked to the mutations in the deletion strains may make it possible to perform recombination experiments on a genome-wide level. The linked nature of the molecular tags may also allow the entire collection of deletion strains to be mated into different genetic backgrounds. However, many applications only require that the molecular tag be in the same cell as the mutation. For example, tags could be introduced on a plasmid or into a

with wild-type growth rates (doubling times of 140 minutes), would be reduced by 30% after 10 population doublings and by over 85% after 60 population doublings. This dramatic change in the relative abundance should be easy to detect using the chip-based hybridization assay. Finally, the collection of individual deletion strains generated by this project would be immediately available to: i) confirm phenotypes observed in the pool; ii) make double mutants; and iii) cross into different genetic backgrounds.

Expanding this approach to the genome-wide level will require a high-throughput strategy to generate the thousands of tagged deletion strains. Towards this goal, a robust PCR deletion strategy has been developed in which all of the steps are either automated or performed in 96-well plates. A 96-well automated multiplex oligonucleotide synthesizer (A.M.O.S.) has also been developed to make the thousands of required oligonucleotides¹². This machine has been designed to minimize the labour and reagent consumption involved in oligonucleotide synthesis. Over 20,000 oligonucleotides have been synthesized on a single machine in one year at a cost of 12 cents per base. The 42,000 oligonucleotides that will be required to delete all of the ORFs in the yeast genome (one 86mer, one 68mer and five 18mers for each of the 6,000 yeast ORFs) could be synthesized in less than a year using the three 96-well synthesizers currently in operation in our laboratory.

Developing a system capable of analysing thousands of tagged deletion strains is also essential to scaling this approach to the genome-wide level. The factors that must be addressed are tag amplification, array density, sensitivity of the detection system and cross-hybridization. The problem of biasing the relative levels of the different tags during the amplification step has been

BEST AVAILABLE COPY

common region of the genome to generate a population of tag-containing wild-type cells. This mixture could be randomly mutagenized and strains with useful phenotypes (such as temperature sensitive mutations) could be identified using the molecular tags. Alternatively, tags could be individually introduced into an existing collection of mutants.

The molecular bar-coding strategy is not limited to analysing deletion strains in yeast. This approach can be applied to any task, *in vitro* or *in vivo*, that requires large populations (cells, DNA fragments, molecules, and so on) to be monitored in parallel. The tags can be introduced in a directed fashion as described, or randomly using either DNA cloning techniques or transposons. Approaches that allow parallel analysis of complex populations will be essential tools as researchers begin to study organisms on the genome-wide level.

Methods

Selection strategy to identify the tag sequences. Starting with all possible 20mers (1.2×10^{12}), a computer algorithm was used to eliminate sequences with known hybridization problems (for example, secondary structure, runs of single nucleotides, unbalanced base composition, and so on). Sequences containing common 9mers (or more) were eliminated during this step because contiguous stretches of homology (≥ 10 bp) have been shown to cause cross-hybridization (data not shown). The output of this programme was a list of 51,082 20mers that are predicted to be as different as possible yet still have similar melting temperatures. For increased specificity, the resulting set of 51,082 20mers was passed through a second filter in which a pair-wise analysis strategy was used to identify and eliminate the sequences most likely to cross-hybridize. Performing the pair-wise analysis using higher thresholds generated sets with fewer but more specific 20mers. Sets containing 51,082, 9,105, 2,643, 853, 170 and 42 20mers were obtained as the stringency level was incrementally increased. The 4,500 20mers described here were selected from the set containing 9,105 20mers.

Hybridization to the high-density oligonucleotide array. Fluorescently labelled oligonucleotides, complementary to 120 different tag sequences, were synthesized using an A.M.O.S. 96-well oligonucleotide synthesizer¹². Yields were quantified by measuring absorbance at 260 nm and the oligonucleotides were used without purification in hybridization assays. The 200- μ l hybridization mixture (6x SSPE-T) contained: 0.9 M NaCl, 60 mM NaH₂PO₄, 6mM EDTA and 0.005% Triton X-100. In addition, the solutions contained 0.5 nM control oligonucleotide (5'-fluorescein-CTGAACGGTAGCATCTTGAC-3'), and 1 nM of each of the 120 fluorescently labelled 20mers. Following a 20 min hybridization at 37 °C, the array was washed 10 times with 1x SSPE-T at 22 °C. The wash solution was removed and the hybridization chamber was re-filled with 6x SSPE-T prior to scanning. Hybridization events were detected by using an argon ion laser to excite the fluorescein; the resulting emission was detected using a photomultiplier tube through a 530 nm bandpass filter (Molecular Dynamics). The entire array was read at a resolution of 11.25 μ m in less than 15 min, generating quantitative signal for each of the 4,500 tag sequences. The collected data was analysed using GeneChip software (Affymetrix).

Construction of the pFA-TAG plasmid. A PCR cloning strategy was used to introduce an 18-bp tag priming site into the multiple cloning region of the pFA6-kanMX2 module¹⁰. A 1.3-kb PCR product was generated using the pFA6-kanMX2 module as template and a pair of primers complementary to the regions flanking the marker (underlined sequences). The tag priming site is included as an 18-base extension on the 3' end

of one of the primers (5'-CTATAGTGTGCTGATTAAAG-CTTCGTACGCTGCAG-3'). The other primer does not have any additional sequence information on the 5' end (5'-GGC-CAGTAGTGGATCTGA-3'). The reaction mixture contained: 20 mM Tris (pH 8.8), 10 mM KCl, 10 mM (NH₄)₂SO₄, 2.0 mM MgCl₂, 0.1% Triton X-100, 250 μ M each dNTP, 0.2 μ M each primer and 2.5 U Vent DNA polymerase (New England Biolabs) in a 100 μ l total volume. The conditions were as follows: 94 °C for 3 min, then 25 cycles of 94 °C for 15 s, 50 °C for 15 s, 72 °C for 30 s, followed by a 3 min extension at 72 °C. The gel-purified PCR product was cloned into *Sma*I linearized pBluescriptSK⁺II (Stratagene) and transformed into electro-competent DH5a cells (GibcoBRL). White colonies were picked from an LB plate (100mg/l ampicillin, IPTG and X-gal) and analysed using PCR. A *kan*^r specific primer (5'-CCTCGACAT-CATCTGCCC-3') along with vector-specific primers (-20 and reverse) were used to confirm the presence and orientation of the insert. The resulting pFA-TAG plasmid was partially sequenced to determine if any mutations were introduced during the cloning procedure.

PCR-targeting strategy. The pFA-TAG module was amplified using a pair of primers that have 3' ends (18 bases) homologous to the marker and 5' ends (30–50 bases) that contain yeast targeting sequences^{3,10}. One of the long oligonucleotides also contained a unique 20mer tag and 18mer common tag priming site (Fig. 2). The *kan*^r gene (aminoglycoside phosphotransferase) from the bacterial transposon Tn903 was used as the dominant selectable marker¹⁰. Expression of the *kan*^r gene confers resistance to the drug G418 in yeast¹⁴. The 100 μ l PCR reaction contained: 2.5 U of Taq DNA polymerase, 10 mM Tris (pH 8.5), 50 mM KCl, 1.5 mM MgCl₂, 0.2 μ M each primer, 250 μ M each dNTP and 10 ng of pFA-TAG template. The reaction conditions were: 3 min at 94 °C, then 30 cycles of 15 s at 94 °C, 15 s at 50 °C and 30 s at 72 °C followed by a 3 min extension at 72 °C. Due to limits in oligonucleotide synthesis chemistry, the proportion of full-length 86mers in an unpurified sample is low. PCR products generated with primers of this length result in low transformation efficiencies because of the missing yeast targeting information on the ends (Fig. 2; blue dashed lines). Purification by HPLC, although suitable, is labour-intensive and impractical to perform on large numbers of oligonucleotides. Instead, the amount of full-length PCR products was increased through a second round of PCR using 20mers that are homologous to the ends of the first PCR product. The second round flanking primer PCR was performed using 1 μ l of the first PCR as template and a pair of 20-base primers that are homologous to the ends of the yeast targeting sequences (Fig. 2). The amplification conditions for the flanking PCR are identical to the conditions used in the first round. The PCR product was analysed on a 1% agarose gel and used without purification in the subsequent transformation.

Transformations. Yeast growth medium and the standard techniques for manipulating yeast have been described¹⁵. The haploid yeast strain YJM826 [*MAT α* *SUC2* *CUP1* Gal⁺], isogenic with S288C, was used for the transformations (provided by John McCusker). Cells (1×10^6) were transformed with approximately 1 μ g of a linear PCR product using a standard lithium acetate method¹⁶. Transformed cells were grown at 30 °C in YPD for 4 h and plated on YPD plates containing 300 mg/l G418 (Gibco BRL). Large G418-resistant colonies (5 to 50) appeared after 3 d growth at 30 °C and the background was low. Colonies (8) from each transformation were streaked onto fresh plates and single colonies were isolated and analysed using PCR with three primers¹¹. In addition, the putative colonies were checked by replica plating onto the appropriate drop-out medium for each of the 11 different auxotrophic ORFs (data not shown). The confirmed deletion strains were stored in 15% glycerol at -70 °C. The 11 auxotrophic deletion strains described in Table 1 were all complete deletions (start codon to stop codon).

BEST AVAILABLE COPY

Generating the deletion pool. Separate 20-ml liquid YPD cultures were inoculated with each of the 11 different auxotrophic tagged deletion strains. The cultures were grown at 30 °C until mid-exponential phase (0.5–1.0 OD₆₀₀). An equal number of cells were combined to generate a master deletion pool which was stored in 10-OD₆₀₀ aliquots in 15% glycerol at –70 °C.

Tag amplification and hybridization. The asymmetric tag amplification reaction was performed on 1 µg of genomic DNA isolated from a mid-exponential phase culture of the deletion pool as previously described¹⁷. One of the common tag amplification primers is 5' labelled with 5(6)-carboxyfluorescein and included in 10-fold excess over the unlabelled primer in this asymmetric PCR. The 100-µl reaction mixture contained: 10 mM Tris (pH 8.5), 50 mM KCl, 1.5 mM MgCl₂, 0.5 µM F-common 1 (5'-fluorescein-GCTTTAATACGACTCACTATAG-3'), 0.05 µM common 2 (5'-GATGTCCACGAGGTCTCT-3'), 250 µM each dNTP, and 2.5 U AmpliTaq DNA polymerase (Perkin-Elmer). The conditions were as follows: 94 °C for 3 min, then 30 cycles of 94 °C for 15 s, 50 °C for 15 s, 72 °C for 30 s, followed by a 3 min extension at 72 °C. The single-stranded, fluorescently labelled 56mer tag amplicons were directly hybridized to the array without purification. The 200-µl hybridization mixture contained: 6x SSPE-T, 0.5 nM control oligonucleotide and 50 µl of the tag amplification reaction (~25 nM mixture of 11 Tags; 2.2 nM each). The hybridization, washing and scanning conditions were as described above.

Deletion pool growth study. An aliquot of the deletion pool was cultivated in synthetic complete (SDC) medium at 30 °C and aliquots containing 3 × 10⁷ cells were harvested from the culture at 0, 3, 6, 9, 12 and 23 h for subsequent analysis on the high-density array. The doubling time for the population is approximately 140 min in SDC medium. To ensure that the cells were harvested from mid-log phase cultures at each of the time points, a series of 2-fold dilutions of the deletion pool were made at the beginning of the growth experiment. Specifically, 10 ml cultures containing 3 × 10⁸, 1.5 × 10⁸, 7.5 × 10⁷, 3.8 × 10⁷, 1.9 × 10⁷, 9.4 × 10⁶, 4.7 × 10⁶ and 2.3 × 10⁶ cells were generated by making 2-fold serial dilutions with SDC medium. At each time point, ~3 × 10⁷ cells were harvested from a culture that was closest to mid-log phase (between 0.5–1.0 OD/ml). This procedure was repeated using SDC-ADE and SDC-TRP medium.

Data analysis. The growth rates of the tagged deletion strains in the pool are measured relative to a reference strain. The reference strain has a tag integrated into a region of the genome that is predicted not to affect the fitness of the cell. At each time point, the signal intensities for each of the elements on the

array were measured. After subtracting the background, the values were normalized to the signal from the reference strain to account for variations in the tag amplification and hybridization conditions. The 'fraction of initial' values were obtained by dividing the normalized values from each time point by the normalized values from the initial time point. A ratio of one indicates that the relative abundance of that deletion strain has not changed in the population. A ratio less than one indicates that the deletion strain was depleted from the population.

In the plot shown in figure 4d (ADE minus medium), the *trp2Δ* deletion strain was used as the reference strain. Similar results were obtained when any of the other non-ADE deletion strains (e.g. *aro2Δ*, *aro7Δ*, *trp3Δ*, *trp4Δ*, and *trp5Δ*) were used as the reference strain. In the future, several different control strains will be constructed and included in the deletion pool. These strains will have tags integrated into regions of the genome that should not affect the fitness of the cell (in pseudo genes, duplicated genes, intergenic regions and so on).

Colony colour assay. An aliquot of the deletion pool was diluted as described above and grown in synthetic complete medium. At various time points (0, 3, 6, 9, 12 and 23 h) several thousands of cells were taken from the pool and plated on rich medium that was not supplemented with adenine. The number of red colonies (*ade1Δ* and *ade2Δ* deletion strains) and white colonies (other nine deletion strains) was determined for each of the time points. The percentage of red colonies was determined at each time point and the values were normalized to the percentage of red colonies that were observed in the starting pool (such as 18%). This process was repeated in medium missing adenine.

Pool size. The concentration of a tag in a hybridization mixture is inversely proportional to the total number of tags in the pool. Tags amplified from a pool containing 6,000 strains would result in an individual tag concentration of 85 pM (100 pmoles can be generated by a single tag amplification reaction/ 6,000 strains = 17 fmoles of each tag which is hybridized in a volume of 200 µl). This concentration is well within the detection limits of the fluorescent scanner.

Acknowledgements

We thank T. Gingeras, S. Fodor, E. Hubble, E. Gentzen, D. Lockhart, R. Sapolski, and M. Chee from Affymetrix; E. Winzler, W. Shih, M. Schena, R. Cho, M. Campbell, J. McCusker, D. Richards, and G. Cavet for helpful discussions and a critical reading of the manuscript.

Received 23 August; accepted 1 November 1996.

1. Oliver, S.G. From DNA sequence to biological function. *Nature* 379, 597–600 (1995).
2. Rothstein, R. Targeting, disruption, replacement, and allele rescue: integrative DNA transformation in yeast. *Meth. Enzymol.* 194, 281–301 (1991).
3. Baudin, A., Ozier-Kalogeropoulos, O., Denouel, A., Lacroute, F. & Cullen, C. A simple and efficient method for direct gene deletion in *Saccharomyces cerevisiae*. *Nucl. Acids Res.* 21, 3329–3330 (1993).
4. Burns, N. et al. Large-scale analysis of gene expression, protein localization, and gene disruption in *Saccharomyces cerevisiae*. *Genes Dev.* 8, 1087–1105 (1994).
5. Smith, V., Botstein, D. & Brown, P.O. Genetic footprinting: a genomic strategy for determining a gene's function given its sequence. *Proc. Natl. Acad. Sci. USA* 92, 6479–6483 (1995).
6. Hensel, M. et al. Simultaneous identification of bacterial virulence genes by negative selection. *Science* 269, 400–403 (1995).
7. Fodor, S.P. et al. Light-directed, spatially addressable parallel chemical synthesis. *Science* 251, 767–773 (1991).
8. Passa, A.C. et al. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc. Natl. Acad. Sci. USA* 91, 5022–5026 (1994).
9. Lockhart, D., J. et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotech.* (in the press).
10. Wach, A., Brachet, A., Pohlmann, R. & Philippsen, P. New heterologous modules for classical or PCR-based gene disruptions in *Saccharomyces cerevisiae*. *Yeast* 10, 1793–1808 (1994).
11. Huxley, C., Green, E.D. & Dunham, I. Rapid assessment of *S. cerevisiae* mating type by PCR. *Trends Genet.* 6, 238 (1990).
12. Lashkari, D.A., Hunnicke-Smith, S.P., Norgren, R.M., Davis, R.W. & Brennan, T. An automated multiplex oligonucleotide synthesizer: development of high-throughput, low-cost DNA synthesis. *Proc. Natl. Acad. Sci. USA* 92, 7912–7915 (1995).
13. Lipshutz, R.J. et al. Using oligonucleotide probe arrays to access genetic diversity. *BioTech.* 19, 442–447 (1995).
14. Jimenez, A. & Davies, J. Expression of transposable antibiotic resistance elements in *Saccharomyces*. *Nature* 287, 869–871 (1980).
15. Sherman, F., Fink, G.R., Hicks, J.B. *Methods in yeast genetics: A laboratory manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1986).
16. Gietz, R.D. & Woods, R.A. High efficiency transformation with *Buham* acetate in *Molecular Genetics of Yeast, A Practical Approach* (ed. Johnston, J.R.) 121–134 (IRL Press, Oxford, UK, 1994).
17. Hoffman, C.S. & Winston, F. A ten-minute DNA preparation from yeast efficiently releases autonomous plasmids for transformation of *Escherichia coli*. *Gene* 57, 287–272 (1987).

BEST AVAILABLE COPY

Multiplexed genotyping with sequence-tagged molecular inversion probes

Paul Hardenbol^{1,3}, Johan Banér², Maneesh Jain^{1,3}, Mats Nilsson², Eugeni A Namsaraev^{1,3}, George A Karlin-Neumann^{1,3}, Hossein Fakhrai-Rad^{1,3}, Mostafa Ronaghi¹, Thomas D Willis^{1,3}, Ulf Landegren² & Ronald W Davis¹

We report on the development of molecular inversion probe (MIP) genotyping, an efficient technology for large-scale single nucleotide polymorphism (SNP) analysis. This technique uses MIPs to produce inverted sequences, which undergo a unimolecular rearrangement and are then amplified by PCR using common primers and analyzed using universal sequence tag DNA microarrays, resulting in highly specific genotyping. With this technology, multiplex analysis of more than 1,000 probes in a single tube can be done using standard laboratory equipment. Genotypes are generated with a high call rate (95%) and high accuracy (>99%) as determined by independent sequencing.

The availability of large collections of SNPs along with recent large-scale linkage disequilibrium mapping efforts¹ have brought the promise of personalized whole-genome association studies to the field of human genetics. To achieve this goal, methodologies that permit screening of hundreds of thousands of SNPs will be needed to implement such large-scale association studies on a routine basis. These methods not only will have to be inexpensive per SNP screened, but will need to consume very little genomic DNA—that is, no more than is typically obtained from a patient's blood sample. In addition, such technologies should ideally require minimal investment in infrastructure so that the technology can be made broadly available.

The challenge of genotyping the approximately 150 molecules of a given SNP locus present in 1 ng of genomic DNA is commonly met by PCR amplification of the locus before genotyping is done². However, an increase in the number of target sequences for simultaneous amplification by PCR quickly leads to unmanageable levels of cross-reaction among primer pairs^{3,4}, whereas parallel hybridization on microarrays^{5,6} lacks the specificity and sensitivity required to genotype large genomes directly.

There are only a limited number of genotyping technologies with sufficient specificity to identify an SNP from genomic DNA without prior PCR amplification. Flap endonucleases have been used to generate a sequence-specific endonuclease cascade in an isothermal fashion that can be assessed with FRET probes^{7,8}. However, this technology is not readily multiplexed for high-throughput applications. Padlock probes are linear oligonucleotides, whose two ends can be joined by ligation when they hybridize to immediately adjacent target sequences⁹. As shown before^{10–12}, padlock probes provide sufficient specificity to analyze SNPs directly, without previous amplification of the target sequences.

Unlike amplification strategies such as PCR and the Invader assay that require two specific primers, cross-reactive padlock probes can easily be distinguished from the desired circular products by methods such as exonucleolysis⁹. This offers the opportunity to add a complex pool of padlock probes to individual DNA samples to investigate large sets of genes in parallel, without a concomitant increase in the risk of cross-reactivity between different probes.

Here we present a strategy that combines DNA detection specificity and sensitivity with the potential to analyze large numbers of target sequences in parallel. Sets of padlock probes with universal tag sequences were reacted with target DNA, molecularly inverted, amplified together and identified in a multiplex analysis yielding more than 1,000 genotypes simultaneously. Using molecular inversion probes, the information content of the SNPs was reformatted into tag sequences that could be detected using a universal oligonucleotide detection array¹³. We report the application of this technique at unprecedented levels of multiplexing, resulting in a lowering of the scale, cost and sample requirements of high-throughput genotyping. The approach retained high accuracy through multiple hybridization and enzymatic processing events, and provided inherent quality control checking.

RESULTS

Selection for circularized probes using exonucleases

Most genotyping methods require PCR amplification of the region spanning the sequence variation. However, when sets of n PCR primer pairs are combined in one reaction to evaluate n target sequences, any of the $2n^2 + n$ possible pairwise primer combinations may give rise to nonspecific amplification products³. With padlock probes the corresponding cross-reactive ligation products create linear dimeric molecules, easily distinguished from circularized

¹Stanford Genome Technology Center, Stanford University, 855 California Avenue, Palo Alto, California 94304, USA. ²The Beijer Laboratory, Department of Genetics and Pathology, Rudbeck Laboratory, Se-751 85 Uppsala, Sweden. ³Present address: ParAllele BioScience 384 Oyster Point Blvd Suite 8, S. San Francisco, California 94080, USA. Correspondence should be addressed to M.R. (mostafa@stanford.edu).

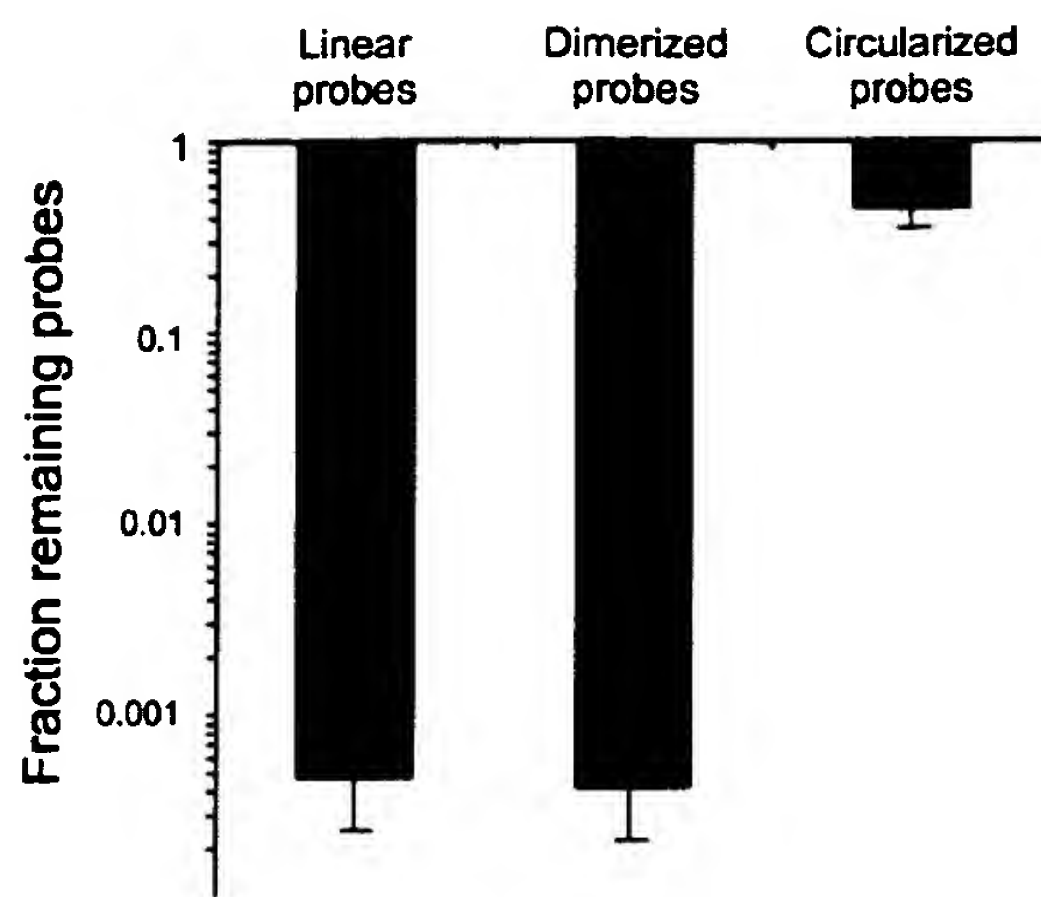


Figure 1 Selection for circularized padlock probes. Effect of exonuclease on linear monomer, dimer or on circularized padlock probes were measured by real-time PCR. Dimerized probes were produced using a ligation template that allowed two different padlock probes to be joined. The results were converted to numbers of molecules by reference to a standard dilution series. The fractions of remaining probe were calculated by dividing each reaction by the respective starting number. Error bars denote s.d. of the ratios from eight reactions.

probes by exonucleolytic degradation^{9,14}. The exonuclease treatment protocol reduces the number of such linear monomeric and dimeric molecules by almost three orders of magnitude with negligible effects on circularized probes as measured by real-time PCR (Fig. 1). The removal of unreacted probes further reduces ligation-independent amplification events that may otherwise occur through accidental priming or templating of polymerization by the large number of linear probes (data not shown).

Molecular inversion probe (MIP) genotyping

Initially we combined pairs of padlock probes specific for alternate alleles in SNP loci. This permitted parallel genotyping of several loci in a single reaction before amplification and identification of the reaction products on tag arrays (Fig. 2b). Before increasing the multiplexing level, we redesigned the padlock probes to be locus-specific to avoid the need for balancing allele-specific probes at every locus (Fig. 2a). With this strategy only one probe was required per locus. To achieve this, the polymorphic nucleotide at the 3' end of the probe was left out, creating a gap between the probe ends. This gap was then filled in four separate allele-specific polymerization (A, C, G and T in four different tubes) and ligation reactions¹⁵. Next, the probes were released from the genomic DNA by removing the uracil residues between primer sequences to avoid topological inhibition of the polymerization reaction¹⁶. The oligonucleotide probe undergoes a unimolecular rearrangement before amplification (Fig. 2b). Each probe contains a unique 20-base tag sequence that is complementary to a sequence on an Affymetrix GenFlex Tag Array. The tags are selected to be similar in melting temperature (T_m) and base composition, and maximally orthogonal in sequence complementarity. These tags amplify and hybridize under a single set of conditions with minimal cross-hybridization to each other and to other features on the microarray.

After amplification, the products are hybridized on four DNA microarrays and the components are decoded by measuring the

fluorescence signals at the corresponding complementary tag site on the DNA array (Fig. 3a). An image of 938 amplified probes hybridized to a DNA array is shown (Fig. 3b). Four intensity values for each probe are generated. The two values for the expected allelic bases are compared to determine whether the sample is homozygous or heterozygous for the given SNP, and the two non-allele bases are compared to the allele bases to determine the signal-to-noise ratio (SNR) for the probe (Table 1). The two non-allele bases serve as internal controls that are used to reduce incorrect genotype calls owing to missing, degraded or noisy probes.

Assay performance

To investigate the performance of the method, probes were generated for 1,121 SNPs from the SNP consortium (TSC) database (<http://snp.cshl.org>) for a 16-megabase region on chromosome 6 centered on the linkage peak for IgA nephropathy¹⁷ (Table 2). Markers were selected from the database based on map position. Of the 1,121 probes, 183 (16%) were inactive during a single synthesis step, possibly owing to such problems as errors in the database, probe design, or failures of oligonucleotide synthesis, probe synthesis or the assay itself. In a pilot study, 25 different individuals were genotyped with the 938 active probes for a total of 23,450 assays. We successfully called 21,336 full genotypes (two chromosomes) and 1,746 half genotypes (single chromosome) (95%) with a median SNR of 16.7 for allele-specific signal to non-allele signal. Half genotypes are reported when the identity of only one of two chromosomes is certain. A cluster plot of data of four of the probes used to genotype 25 individuals is shown (Fig. 4). No substantial differences were seen in the call rates of probes designed for all allele combinations (Table 3).

Accuracy was determined through independent sequencing. 1,517 loci were genotyped in a 1,517-probe multiplex analysis with ten individuals. Forward and reverse Sanger sequencing was performed on a subset (129) of PCR amplicons of 1,517 loci amplified from the same 10 individuals. Conservative reads were made manually with the identity of the forward and reverse loci blinded at the time of sequence interpretation. Accuracy of Sanger sequencing was measured by comparing reads for which the sequence of both strands existed. 359 of 367 sequence pairs were identical, for an

Table 1 Data generated from the first 10 probes from individual NA17203

| Probe ID | Allele | Base call ^a | Signal A ^b | Signal G | Signal C | Signal T | SNR ^c |
|----------|--------|------------------------|-----------------------|----------|----------|----------|------------------|
| 2,515 | A/G | G/G | 139 | 1,472 | 216 | 202 | 6.8 |
| 2,516 | A/G | A/A | 437 | 21 | 30 | 31 | 14.1 |
| 2,517 | A/G | A/G | 1,538 | 1,494 | 95 | 94 | 16.2 |
| 2,518 | A/G | A/G | 343 | 474 | 39 | 30 | 12 |
| 2,519 | A/G | A/A | 3,574 | 39 | 51 | 65 | 55.2 |
| 2,520 | A/G | G/G | 147 | 1,702 | 175 | 172 | 9.8 |
| 2,521 | A/G | G/G | 59 | 1,290 | 45 | 38 | 28.5 |
| 2,522 | A/G | A/G | 478 | 382 | 110 | 87 | 4.4 |
| 2,523 | A/G | G/G | 36 | 1,234 | 49 | 62 | 19.9 |
| 2,524 | A/G | G/G | 62 | 1,492 | 59 | 115 | 13 |

^aA base-call is made if the SNR is at least 3, and the ratio of the higher allele signal to the lower allele signal is >6:1 for homozygous calls and <2:1 for heterozygous calls. ^bSignals are normalized fluorescence intensity values produced by the microarray scanner. ^cThe SNR was calculated for each marker by dividing the maximum allele signal (A and G) by the maximum signal in the two background channels (C and T).

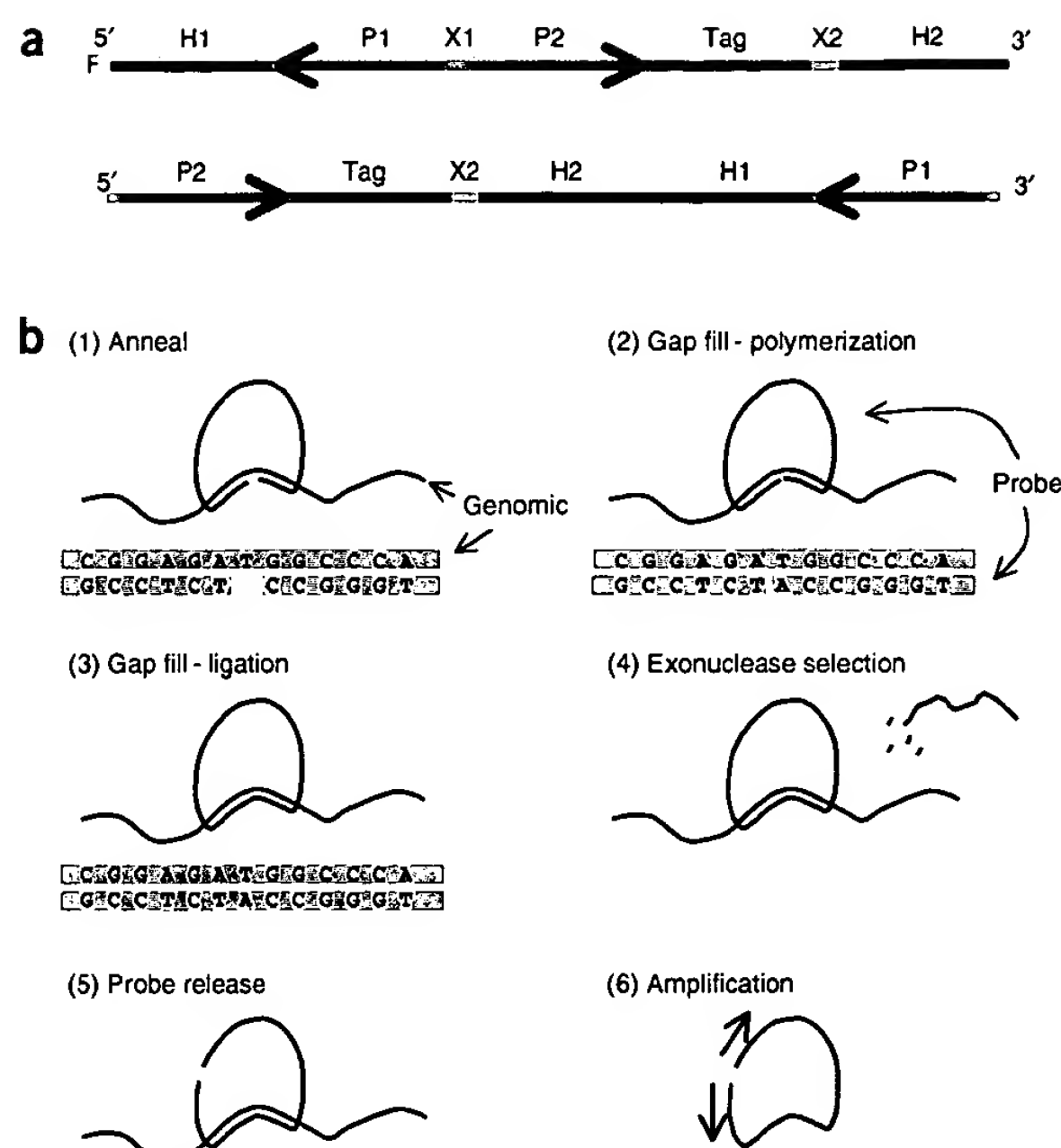


Figure 2 Molecular inversion probes. (a) Unreacted probe (top) and inverted probe (bottom). A single probe is used to detect both alleles of each SNP and consists of seven segments: two regions of homology to target genomic DNA, H1 and H2 (unique to each probe) at the termini of the probe, two PCR primer regions common to all probes, one bar code specific for each locus and two common cleavage sites, X1 and X2. Successfully reacted probes are amplified using primers P1 and P2. A universal detection tag sequence, one of 16,000, is for array detection of amplified probe. Cleavage sites X1 and X2 are used to release the circularized probe from genomic DNA and for post-amplification processing, respectively. (b) Enzymatic probe inversion. (1) A mixture of Genomic DNA, 1,000 or more probes, and thermostable ligase and polymerase is heat-denatured and brought to annealing temperature. Two sequences targeting each terminus of the probe hybridize to complementary sites in the genome, creating a circular conformation with a single-nucleotide gap between the termini of the probe. (2) Unlabeled dATP, dCTP, dGTP or dTTP, respectively, is added to each of the four reactions. In reactions where the added nucleotide is complementary to the single-base gap, DNA polymerase adds the nucleotide and (3) DNA ligase closes the gap to form a covalently closed circular molecule that encircles the genomic strand to which it is hybridized. (4) Exonucleases are added to digest linear probes in reactions where the added nucleotide was not complementary to the gap and excess linear probe in reactions where circular molecules were formed. The reactions are then heated to inactivate the exonucleases. (5) To release probes from genomic DNA, uracil-*N*-glycosylase is added to depurinate the uracil residues in the probes. The mixture is then heated to cleave the molecule at the abasic site and release it from genomic DNA. (6) PCR reagents are added, including a primer pair common to all probes. The reactions are then subjected to thermal cycling, with the result that only probes circularized in the allele-specific gap-fill reaction are amplified.

accuracy of 97.8%. The accuracy of the 359 agreeing sequence pairs was therefore 99.95%. In the genotyping data set, 312 full genotypes and 23 half genotypes were identical with the Sanger sequence pairs (643/647 chromosome comparisons), for an accuracy of 99.4%. Similar accuracy was achieved when the data were compared with pyrosequencing^{18,19} data on different sets of SNPs. The four discordant genotypes were found in a single locus (probe 2101) in four individuals. Because the sequencing data were nonpolymorphic and subsequent sequencing of these loci with newly generated amplicons was concordant with our genotyping results, it is likely that the original amplification for Sanger sequencing was contaminated with other template DNA or PCR product from another individual.

To measure the reproducibility of MIP genotyping, an individual was assayed three times with 888 markers (5,328 chromosomes).

Table 2 Performance metrics of the genotyping assay by molecular inversion probes

| Measured parameter | Result |
|---|--------|
| Design rate ^a | 84% |
| Call Rate ^b | 95% |
| Concordance with Sanger sequencing ^c | >99.4% |
| Repeatability ^d | 99.9% |
| Highest multiplex level | 1,517 |
| Average SNR ^e | 17 |
| Genomic DNA used / SNP ^f | 2 ng |

^a183 of 1,121 probes failed to generate data. ^bAn average of 891 of 938 probes called per individual for 25 experiments. ^cTwo of 396 chromosomes were discordant with pyrosequencing. ^d5,006 of 5,011 chromosome comparisons were concordant. ^eAverage of the ratio of maximum allele signal to maximum non-allele signal of called probes. ^f2 µg genomic DNA used to genotype 1,121 markers per individual.

5,006 of 5,011 chromosome comparisons were concordant (99.9%) (Table 2).

We investigated the effect of increasing the multiplexing level tenfold. The performance of 75 probes either in a 75-probe multiplexed reaction or embedded in a 938-probe multiplex reaction was compared on the same individual's DNA (Table 3). The average call rate in seven repetitions of the same individual for the 75-probe multiplex was 92.6%. Call rate for the same 75 probes in the 938-probe multiplex was 93.4% (average of 25 individuals). The assay conditions were identical in every respect except the number of probes added.

Because DNA array costs represent a substantial fraction of the overall cost of this method, we compared four-chip-one-color detection to two-chip-two-color detection in otherwise identical experiments. The dyes were carboxyfluorescein directly coupled to the labeling oligonucleotide, and phycoerythrin that was coupled to the labeling oligonucleotide via biotin-streptavidin in post-chip hybridization staining²⁰. Call rate and SNR in the two-chip-two-color experiment (96.1% and 30, respectively) were very similar to those in the four-chip-one-color experiment (95.8% and 31).

DISCUSSION

The MIP genotyping method described here has several advantages over alternative techniques. No singleplex PCR amplification is required before mutation detection, thereby reducing labor and expense. PCR is applied only after mutation detection, at which time all molecular inversion probes are converted to standard-length oligonucleotides of similar sequence composition and common primers. This results in a high degree of multiplexing capacity. We have not observed any change in performance in multiplexing from a single probe up to 1,500 probes and speculate that a further increase to 10,000 probes might be possible because sufficient signal is generated in the assay to support that many probes. The data

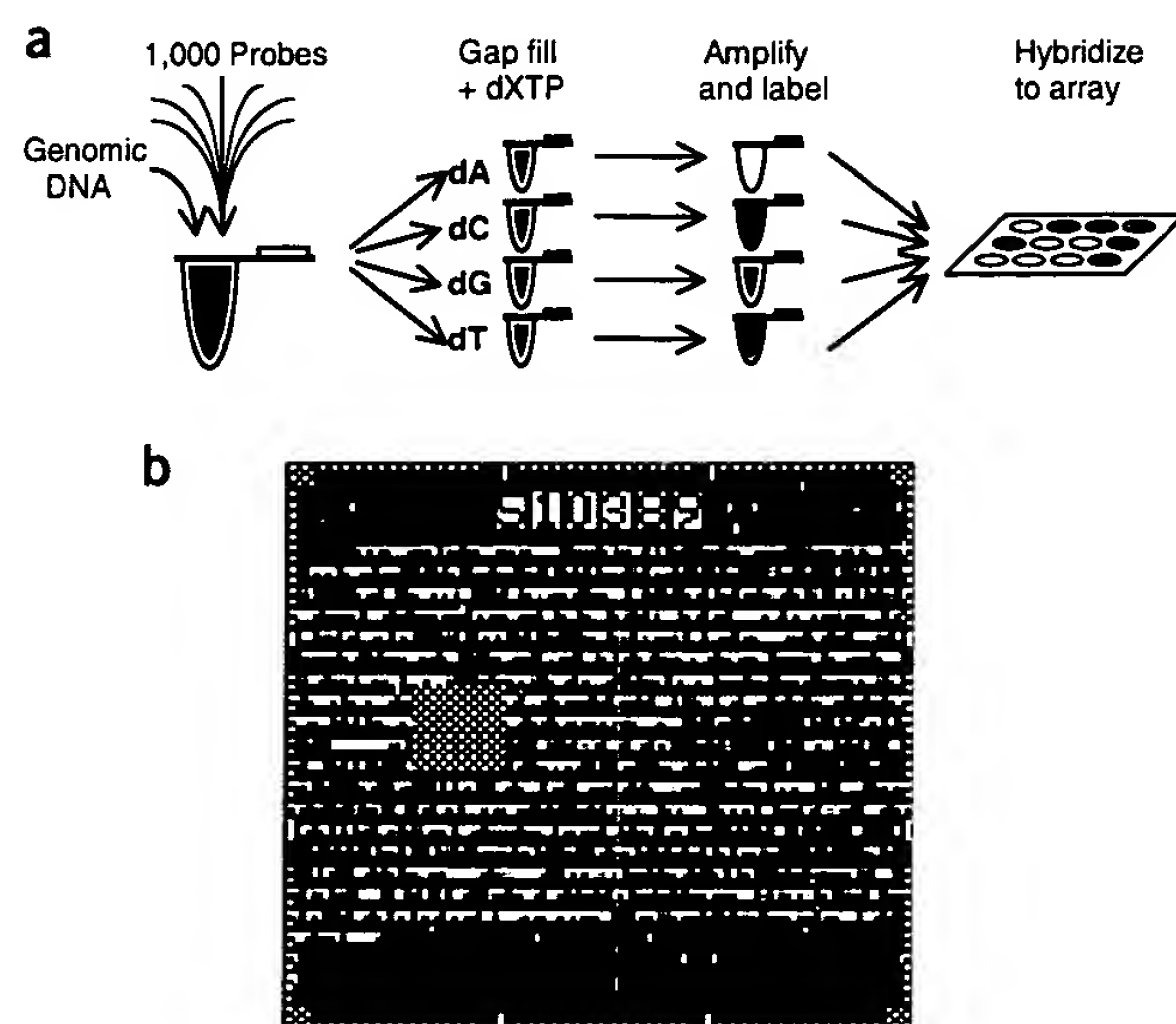


Figure 3 Process flow and array image. (a) Genotyping process flow. 1,000 or more probes are mixed with genomic DNA and gap-fill enzymes (see Fig. 2). The reaction is split into four tubes and one of four unmodified nucleotides is added. Reactions are subsequently amplified and a label is added. Reactions are combined and hybridized to the microarray. Relative intensities of two expected allele bases and two background bases indicate genotype and probe performance. (b) Data from 938 amplified probes hybridized to a GenFlex universal DNA array. The relative base incorporation is measured by the fluorescence signals at the corresponding complementary tag site on the DNA array.

presented here were generated using four microarrays per sample. Currently we use two microarrays with two-color detection per sample as previously described²⁰, and we obtain equivalent call rates and SNRs. In theory, genotyping 16,000 markers with this method would require 44 reactions and 2 oligonucleotide arrays (1,500-probe multiplex with 16,000 element Affymetrix Tag 3 array using two-color detection). Thus only a very modest infrastructure is needed to use this approach: a small number of thermocyclers, microarray washing instruments and microarray scanners. This compares very favorably with the robotic infrastructure and detection instrumentation required to set up thousands of PCR reactions and analyze the results.

The intramolecular nature of the MIP genotyping allows higher multiplexing than any other current approach because only the self-self interacting molecules are amplified, while cross-interactions are greatly suppressed. This should allow the current DNA usage of 2 ng per SNP reaction (2 μ g/1,000 probes) to be further reduced to 0.2 ng per SNP reaction (2 μ g/10,000 probes) as the degree of multiplexing is increased to 10,000 probes.

Several levels of intrinsic specificity are built into this assay. First, the dual recognition sequences at the 3' and 5' ends of probes are physically constrained to interact locally. A molecular inversion probe hybridizes much more quickly than two

independent probes because the second recognition sequence hybridizes instantaneously after the first. As a result, probe-genomic complexes form at probe concentrations that do not favor nonspecific cross-interactions between probes. Specificity is then increased by the action of the gap-fill enzymes. DNA polymerase selectively extends the correct nucleotide, and DNA ligase ligates only perfectly hybridized DNA. An error requires both misextension and misligation to occur. Probes that have undergone the correct interaction and circle formation are further selected by exonuclease treatment before amplification. Finally, the tag sequences are selected to achieve high hybridization specificity and thereby eliminate cross-talk at the detection step. The synergism of the individually optimized steps comprising the MIP genotyping results in the high degree of multiplexing described here.

An unusual aspect of the approach is the built-in quality control of SNR through monitoring of the background allele channels. Bi-allelic markers such as SNPs have only two possible base alleles. Because this assay monitors all four base possibilities, the SNR is measured with each call and suspect calls can be efficiently discarded.

Molecular inversion requires a single probe per marker, reducing the requirement for probe synthesis. Moreover, any damage or loss of performance of that probe will affect both alleles equally and will therefore not lead to spurious genotypes such as can occur with allele-specific oligonucleotides.

For molecular inversion technology, the rate at which a functional probe is generated from an SNP chosen at random from a database in a single synthesis attempt is 84%. The rate at which all functional probes produce high-quality data over many individuals is 95%. As mapping and cSNP (SNPs that are found in exons) discovery efforts proceed, it will be increasingly important to assay a particular SNP rather than any SNP within a region. This will place increasing emphasis on the ability of a given technology to assay any SNP.

Cost is a fundamental driver for the development of alternative SNP genotyping technologies. There are three main costs associated with SNP genotyping: probe cost, assay cost and detection cost. Although molecular inversion probes are longer than PCR primers, the total number of unique bases that must be synthesized for each probe is comparable to that for a PCR-based approach and much lower than for methods that require allele-specific oligonucleotides, such as the oligonucleotide ligation assay²¹. The locus-specific probes do not require any fluorescent or modified bases and are

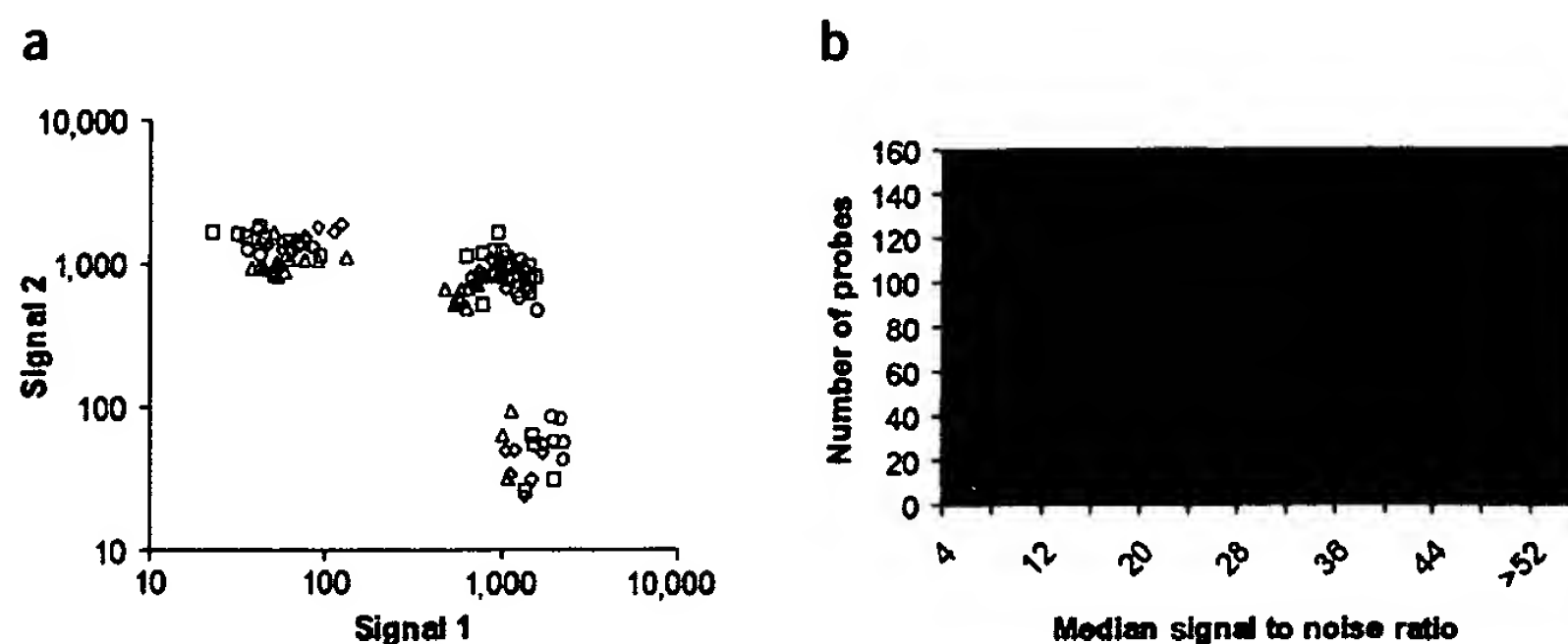


Figure 4 Assay performance. (a) Fluorescence signal from four markers tested on 25 individuals in 32 experiments in a 1,121-probe multiplex assay. Markers 1, 2 and 3 are A/G alleles and marker 4 is a C/G allele. A and C signal is plotted as signal 1 and G signal as signal 2. (b) Median ratio of maximum allele signal to maximum background (non-allele) signal for 938 probes.

Table 3 Data on allele type, median call rate and SNR obtained by large-scale genotyping

| Allele | No. tested | Average call rate | Median SNR |
|--------|------------|-------------------|------------|
| A/C | 77 | 94% | 19 |
| A/G | 349 | 95% | 18 |
| A/T | 52 | 94% | 22 |
| C/G | 90 | 94% | 19 |
| C/T | 302 | 95% | 22 |
| G/T | 68 | 94% | 21 |

therefore inexpensive to synthesize. Also, owing to the kinetic advantage of intramolecular interactions, only 12 amol of each probe are used in a single assay. Typical synthesis scales of 1 nmol thus represent millions of assays worth of material. These probes will thus persist as a valuable resource for subsequent genotyping. Assay costs are amortized by the high degree of multiplexing involved, resulting in a very inexpensive assay in the current format (<\$0.01 per assay). Detection costs are dominated by the cost of the microarray. To minimize this cost, arrays should be fully occupied and larger arrays used to amortize the cost of the arrays over more genotypes. Such arrays are now available from multiple suppliers either on or off the shelf (GenFlex and Tag3 arrays; Affymetrix) or on a service basis (Agilent, Amersham, NimbleGen), and the cost per array feature ranges from less than \$0.01 to \$0.10. Taken together, we believe these advantages amount to an enabling advance in genotyping technology that will allow the power of whole-genome analysis to be realized.

METHODS

Oligonucleotide synthesis. Example of a probe constructed for NCBI SNPcluster ID: rs 1389629 (TSC0559094) 5'-TGGATCCCATTATCCTCCAT-TACGGCTCAACGTTCTTATTCGGTTUUTTCAAATGTTATCGAGGTC-CGGCACGCACAGGTTATGAATCTCTTTAACTCCCACAGTGAGGAGC-3'. Molecular inversion probes were ordered from ParAllele BioScience and Eurogentec. Other oligonucleotides were synthesized by Operon, Inc.

Samples. All samples were obtained from the Coriell Institute human variation collection of African American (HD100AA) and Caucasian (HD100CAU) genomic DNA.

Exonucleolysis assay. Circularization of 3 fmol MIP probe A (5'-TGATGGACGTCTGGAAAGCAACCGAAGCTTGTGCGCGTGACAGGGT-CACCAGCAGGCATGAGCCCGGTCAACTTCAAGCTCCTAAGCCG-GCAGGCAATGCACAGCACCG-3') was done using as template 6 fmol of the oligonucleotide target 5'-TGCTTTCCAGACGTCCATCACGGTGCTGT-GCATTGCCTGC-3', in 20 mM Tris-HCl pH 8.5, 100 mM KCl, 9 mM MgCl₂, 1 mM dithiothreitol, 0.1% Triton X-100 and 4 units Ampligase (Epicentre). Reactions were thermally cycled ten times between 95 °C for 30 s and 58 °C for 1 min. Exonucleolysis was performed by addition of 8 units each of exonuclease I and exonuclease III (New England Biolabs) and incubation at 37 °C for 3 h, followed by denaturation at 95 °C for 10 min. Linear or circularized forms of A were amplified using primers F1 (5'-TGATGGACGTCTG-GAAAGCAA-3') and R1 (5'-CGGTGCTGTGCATTGCCTGCC-3'), or F2 (5'-CACGCGCACAAGCTTCGG-3') and R2 (5'-CAGGGCACCAGCAG-GCA-3'), respectively. Dimerization of the two padlock probes A (sequence above) and B (5'-TGTTCACTGCTGGCCTCCGCAAGCGTGTAGTGTC-CGTCGAATAT-TCGTTCTGCAGCATCGCACACAGAAGGTCGATTGC-TAGGTGACTGCCACCAAGGGG-3') was done using the template oligonucleotide 5'-GCGGAGGCCAGCAGTGAACACGGTGCTGCATTGC-CTGC-3', and the ligation product was amplified with primers F2 and R2. Real-time amplification was done with an ABI Prism 7700 thermocycler

(Applied Biosystems) in a total volume of 30 µl containing 2.5 µl buffer A (Applied Biosystems), 2 mM MgCl₂, 0.2 mM of each dNTP, 0.3 µM of each primer, 0.8 µM TaqMan probe 5'-FAM-CCCGGTCAACTTCAAGCTCC-TAAGCC-TAMRA-3' and 0.6 units AmpliTaq Gold (Applied Biosystems). The temperature profile was 95 °C for 10 min followed by 50 cycles of denaturation at 95 °C for 30 s and annealing and extension at 58 °C for 45 s.

Molecular inversion probe assay. Denaturing and annealing: four identical reactions containing 400 ng of genomic DNA, 12 amol each of 1,000 probes, 0.0625 units Ampligase (Epicentre) and 0.5 units Stoffel fragment DNA polymerase (Applied Biosystems) in 9 µl of 20 mM Tris-HCl (pH 8.3), 25 mM KCl, 10 mM MgCl₂, 0.5 mM NAD and 0.01 % Triton X-100 were incubated for 4 min at 20 °C, 5 min at 95 °C and 15 min at 60 °C. Gap-fill reaction: 1 µl of each of four nucleotides was added to the four reactions and incubated for 10 min at 60 °C and then 1 min at 37 °C. Exonuclease selection: 10 units exonuclease I and 200 units exonuclease III (United States Biochemical) in a 2-µl volume were added and the mixture incubated for 14 min at 37 °C, 2 min at 95 °C and 1 min at 37 °C. Uracil depurination and cleavage: 2 units of uracil-N-glycosylase (New England Biolabs) were added in 25 µl of 1.6 mM MgCl₂, 10 mM Tris-HCl (pH 8.3), 50 mM KCl and incubated for 9 min at 37 °C and 20 min at 95 °C. Amplification: 2 units of AmpliTaq Gold (Applied Biosystems), 16 pmol primer 1 (5'-CCGAATAG-GAACGTTGAGCCGT-3'), and 16 pmol primer 2 (5'-GCAAATGTTATC-GAGGTCCGGC-3') in 25 µl of 1.6 mM MgCl₂, 10 mM Tris-HCl (pH 8.3), 50 mM KCl and 112 µM dNTP were preactivated for 10 min at 95 °C and then added to the genotyping reactions. The reactions were amplified in 28 cycles of 95 °C for 20 s, 65 °C for 45 s and 72 °C for 10 s. Sample processing: 20 units of exonuclease I and 10 units *DraI* (New England Biolabs) were incubated with 60 µl of each amplification product at 37 °C for 1 h and then 80 °C for 30 min.

Microarray hybridization. Approximately 1.25 pmol of each amplified and processed product were hybridized overnight at 39 °C to a GenFlex Tag Array (Affymetrix) DNA array with 55 µl 2× MES, 2.2 µl 50× Denhardt buffer, 110 pmol primer 1 complement (5'-ACGGCTCAACGTTCC-TATTCGG-3'), 110 pmol primer 2-FAM (5'-FAM-GCAAATGTTATCGAG-GTCCGGC-3'), 1.1 fmol (each) GenFlex control oligonucleotide (Affymetrix). Microarrays were washed and scanned as recommended by the manufacturer (Affymetrix). Data analysis was performed on the raw signal data for each tag feature generated by the Affymetrix image analysis software.

Data analysis. Four signals were generated for each genotype, one for each base reaction. The raw signal was background subtracted and then normalized using the GenFlex hybridization controls that were hybridized in equimolar amounts to each microarray. Base-calls were performed as follows: a small systematic noise was added to all four channels (A,C,G,T). Measured signals were adjusted as follows: $S_{\text{adjusted}} = \sqrt{(S_{\text{measured}}^2 + (7 \times \sigma)^2)}$, where σ = 3 times the standard deviation of features that were not used in a particular experiment (array noise). The SNR was calculated for each marker by dividing the maximum signal in the two expected allele channels by the maximum signal in the two background channels. The two criteria for calling genotypes were a signal to background ratio of at least 3:1, and the ratio of the higher allele signal to the lower allele signal greater than 6:1 for homozygous calls and less than 2:1 for heterozygous calls.

ACKNOWLEDGEMENTS

We thank K. Tran, T. Brundage, J. Patterson, E. Prakash and C. Bruckner for technical assistance and T. Jones and P. Oefner for helpful discussions. This work was supported by NIH HG00205. The work in Uppsala was supported by the Beijer and Wallenberg Foundations, the Research Councils of Sweden for natural science and for medicine, the Swedish Cancer Fund and Polysaccharide Research AB (Uppsala).

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests (see the *Nature Biotechnology* website for details).

Received 18 October 2002; accepted 21 February 2003

Published online 5 May 2003; doi:10.1038/nbt821

1. Patil, N. *et al.* Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723 (2001).
2. Syvanen, A.C. Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nat. Rev. Genet.* **2**, 930–942 (2001).
3. Nilsson, M. *et al.* Making ends meet in genetic analysis using padlock probes. *Hum. Mutat.* **19**, 410–415 (2002).
4. Elnifro, E.M., Ashshi, A.M., Cooper, R.J. & Klapper, P.E. Multiplex PCR: optimization and application in diagnostic virology. *Clin. Microbiol. Rev.* **13**, 559–570 (2000).
5. Lockhart, D.J. *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* **14**, 1675–1680 (1996).
6. Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
7. Lyamichev, V. *et al.* Polymorphism identification and quantitative detection of genomic DNA by invasive cleavage of oligonucleotide probes. *Nat. Biotechnol.* **17**, 292–296 (1999).
8. Eis, P.S. *et al.* An invasive cleavage assay for direct quantitation of specific RNAs. *Nat. Biotechnol.* **19**, 673–676 (2001).
9. Nilsson, M. *et al.* Padlock probes: circularizing oligonucleotides for localized DNA detection. *Science* **265**, 2085–2088 (1994).
10. Lizardi, P.M. *et al.* Mutation detection and single-molecule counting using isothermal rolling-circle amplification. *Nat. Genet.* **19**, 225–232 (1998).
11. Faruqi, A.F. *et al.* High-throughput genotyping of single nucleotide polymorphisms with rolling circle amplification. *BMC Genomics* **2**, 4 (2001).
12. Antson, D.O., Isaksson, A., Landegren, U. & Nilsson, M. PCR-generated padlock probes detect single nucleotide variation in genomic DNA. *Nucleic Acids Res.* **28**, E58 (2000).
13. Shoemaker, D.D., Lashkari, D.A., Morris, D., Mittmann, M. & Davis, R.W. Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nat. Genet.* **14**, 450–456 (1996).
14. Zhang, D.Y., Brandwein, M., Hsuih, T.C. & Li, H. Amplification of target-specific, ligation-dependent circular probe. *Gene* **211**, 277–285 (1998).
15. Abravaya, K., Carrino, J.J., Muldoon, S. & Lee, H.H. Detection of point mutations with a modified ligase chain reaction (Gap-LCR). *Nucleic Acids Res.* **23**, 675–682 (1995).
16. Banér, J., Nilsson, M., Mendel-Hartvig, M. & Landegren, U. Signal amplification of padlock probes by rolling circle replication. *Nucleic Acids Res.* **26**, 5073–5078 (1998).
17. Gharavi, A.G. *et al.* IgA nephropathy, the most common cause of glomerulonephritis, is linked to 6q22–23. *Nat. Genet.* **26**, 354–357 (2000).
18. Fakhrai-Rad, H., Pourmand, N. & Ronaghi, M. Pyrosequencing: an accurate detection platform for single nucleotide polymorphisms. *Hum. Mutat.* **19**, 479–485 (2002).
19. Ronaghi, M. Pyrosequencing sheds light on DNA sequencing. *Genome Res.* **11**, 3–11 (2001).
20. Fan, J.B. *et al.* Parallel genotyping of human SNPs using generic high-density oligonucleotide tag arrays. *Genome Res.* **10**, 853–860 (2000).
21. Landegren, U., Kaiser, R., Sanders, J. & Hood, L. A ligase-mediated gene detection technique. *Science* **241**, 1077–1080 (1988).